UNIVERSIDADE ESTADUAL PAULISTA JÚLIO MESQUITA FILHO
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

LUCIANA CANDIDA DA SILVA

Publicação de Dados de Pesquisa Científica: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science

Orientador: Professor Dr. José Eduardo Santarem Segundo

Marília, SP
2020

UNIVERSIDADE ESTADUAL PAULISTA JÚLIO MESQUITA FILHO
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

LUCIANA CANDIDA DA SILVA

Publicação de Dados de Pesquisa Científica: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio Mesquita Filho como parte dos requisitos para obtenção do título de Doutora em Ciência da Informação.

Orientador: Professor Dr. José Eduardo Santarem Segundo

Área de concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Marília, SP
2020

Catalogação na Publicação

Silva, Luciana Candida da.

S856p Publicação de dados de pesquisa científica [manuscrito]: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science / Luciana Candida da Silva. – Marília, 2020.
243 f. : il. ; 30 cm.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP), 2020.
Orientador: Prof. Dr. José Eduardo Santarem Segundo.

1. Dados de pesquisa científica. 2. Cadernos abertos de pesquisa. 3. Web Semântica. 4. Linked Data. 5. e-Science. I. Santarem Segundo, José Eduardo (orientador). II. Título.

CDU: 004.65
CDD: 005.73

Luciana Candida da Silva
Bibliotecária CRB-1 /1831

Como citar esse documento:

SILVA, Luciana Candida da. **Publicação de dados de pesquisa científica**: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science. Orientador: José Eduardo Santarem Segundo. 2020. 243 f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP), Marília, SP, 2020.

LUCIANA CANDIDA DA SILVA

Publicação de Dados de Pesquisa Científica: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio Mesquita Filho (UNESP), como requisito parcial para obtenção do título de doutora em Ciência da Informação.

BANCA EXAMINADORA:

Prof. Dr. José Eduardo Santarem Segundo
Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP)
Universidade Estadual de São Paulo (USP)

Profa. Dra. Silvana Aparecida Borsetti Vidotti
Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP)

Prof. Dr. Leonardo Castro Botega
Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP)

Prof. Dr. Dalton Lopes Martins
Universidade de Brasília (PPGCI/UNB)

Profa. Dra. Liliana Giusti Serra
Software Sophia

Marília-SP, 18 de setembro de 2020.

Dedico esta conquista às minhas estrelas preferidas que acompanharam o desenvolvimento deste estudo lá do céu, José Francisco da Silva e Manoel Francisco da Silva, irmão e pai, e indiscutivelmente, à joia mais preciosa do meu mundo, Júlia Candida da Silva, minha amada mãe, que partiu para o encontro de Deus no transcurso desta tese.

AGRADECIMENTOS

Ao meu professor e orientador, José Eduardo Santarem Segundo, pelo empenho em esclarecer e nortear o desenvolvimento desta tese, pela paciência, cuidado e confiança.

Aos professores membros da banca, Silvana Aparecida Borsetti Vidotti, Leonardo Castro Botega, Dalton Lopes Martins e Liliana Giusti Serra pelas ricas contribuições a este estudo.

Aos professores da linha de pesquisa Informação e Tecnologia do Programa de Pós-Graduação em Ciência da Informação da UNESP, em especial aos professores Eduardo Santarem, Zaira Zafalon, Plácida L. Santos, Silvana Vidotti, Leonardo Botega, José Augusto e Daniel Martinez pelas contribuições em minha formação acadêmica.

À Universidade Federal de Goiás, pela licença concedida para realizar e viver esse grandioso mundo do doutorado. Aos colegas do curso de Biblioteconomia que, mesmo após o fim da licença capacitação, proporcionaram condições para que eu pudesse concluir a tese.

À minha família, em especial aos meus pais e irmãos, os quais me criaram com amor e dedicação, pelas lições de vida, pelas alegrias compartilhadas e por todo aprendizado ensinado. Agradecimento especial à sobrinha Adrielle Cristina por compartilhar os conhecimentos às pesquisas experimentais. Também, ao *brother in law* Seth Hammer Alalof, pelas traduções dos meus textos, e ao Renan pelo apoio e companheirismo.

Aos amigos que direta ou indiretamente contribuíram para a realização desta pesquisa.

À Deus, pela saúde, força e sabedoria!

“Nenhum de nós é tão inteligente como todos nós”

Philip Condit

RESUMO

Vivencia-se um período de mudanças nas práticas científicas, exigindo novas maneiras de gerar e comunicar a ciência. Essa nova maneira implica em disponibilizar dados de pesquisa científica gerados em laboratórios de pesquisa em tempo real, ou o mais próximo disso, em formatos abertos e estrutura adequada para permitir que sejam acessíveis, compartilháveis e reutilizáveis. Neste contexto, vislumbra-se na Web Semântica e no *Linked Data* conceitos e tecnologias que enfatizam a reutilização e a ligação de recursos ricamente descritos na Web. O objetivo geral desta tese é propor diretrizes semânticas para estruturação e publicação de dados abertos de cadernos de pesquisa, visando melhorias na qualidade da sua recuperação e compartilhamento em plataformas de acesso aberto. Nesse sentido, para realizar esse estudo, foram identificados os elementos conceituais e práticos presentes nas dimensões da *e-Science*, e apresentadas as características e as especificidades dos dados científicos anotados em cadernos de pesquisa. Na sequência, descreveu-se os conceitos e tecnologias da Web Semântica e *Linked Data* apropriadas para publicações desses dados em plataformas de acesso aberto. As diretrizes propostas nesta tese adotaram a etapa correspondente à formalização, estrutura, formatos e licenças de Santarem Segundo (2018). A metodologia seguiu os procedimentos tradicionais para delimitação do universo e amostragem da pesquisa, como sua classificação e coleta de dados; e, revisão sistemática de literatura para identificar trabalhos relacionados ao estado da arte dos cadernos abertos de pesquisa, no que se refere a sua estrutura e publicação para acesso e uso dos dados. Definiu-se que a pesquisa é de natureza qualitativa e finalidade aplicada; o método é bibliográfico, descritivo, exploratório, documental e de levantamento. Para a composição das diretrizes semânticas, identificou-se o ecossistema da pesquisa cientifica em torno do caderno de laboratório, realizou a modelagem dos dados a partir do modelo conceitual IFLA LRM, o mapeamento e a definição dos metadados apropriados ao contexto dos cadernos de pesquisa. Logo, os vocabulários selecionados foram descritos, bem como foram indicados os vocabulários para enriquecimento e as licenças de uso. Depois disso, analisou-se as correspondências entre os metadados e as propriedades dos vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets*. Em seguida foi construído o mapeamento das propriedades para relacionamentos de dados. Após o estudo das etapas, analisou-se os elementos quanto ao alcance dos Princípios FAIR e melhores práticas do W3C. Como resultado da pesquisa, estabeleceu-se um conjunto de diretrizes semânticas compostas de elementos e tecnologias que refletem a realidade de pesquisas laboratoriais e a descrição de experimentos com uma pluralidade de atributos, precisos e relevantes, os quais poderão proporcionar benefícios à comunidade cientifica com dados organizados, padronizados e disponíveis para o reuso. A aplicação devida dessas diretrizes, no que se refere à estruturação de dados, colabora para que os dados sejam encontráveis, acessíveis, interoperáveis e reutilizáveis.

Palavras-chave: Dados de pesquisa científica. Cadernos abertos de pesquisa. Cadernos de laboratório. *Linked Data*. Web Semântica. *e-Science*. Diretrizes semânticas.

ABSTRACT

We are in a period of changes in scientific practices requiring new ways to generate and communicate science. This new way implies making available scientific research data generated in research laboratories in real time, or the closest, in open formats and adequate structure to allow the data to be accessible, shareable and reusable. In this context, one sees in the Semantic Web and Linked Data concepts and technologies that emphasize the reuse and connection of resources richly described on the Web. The general objective of this thesis is to propose semantic guidelines for structuring and publishing open data from research notebooks, aiming to improve the quality of data recovery and sharing on open access platforms. For this, the conceptual and practical elements present in the e-Science dimensions were identified. It presented the characteristics and specificities of the scientific data noted in research notebooks. It then described the concepts and technologies of the Semantic Web and Linked Data appropriate for publishing these data on open access platforms. The guidelines proposed in this thesis adopted the stage corresponding to the formalization, structure, formats and licenses of Santarem Segundo (2018). The constructed methodology followed the traditional methodological procedures for delimiting the universe and sampling the research, classification of the research and data collection; and, systematic literature review to identify works related to the state of the art of open notebooks science with regard to their structure and publication for access and use of data. It was defined that the research is of a qualitative nature and applied purpose; the method is bibliographic, descriptive, exploratory, documentary and survey. For the composition of the semantic guidelines, the ecosystem of scientific research was identified around the laboratory notebook, performed data modeling based on the IFLA LRM conceptual model, mapped and defined the metadata appropriate to the context of the research notebooks. It then described the selected vocabularies, as well as indicated vocabularies for data enrichment and usage licenses. After that, it analyzed the correspondences between the metadata and the properties of the Schema.org, DC Terms, SKOS and RDA Element Sets vocabularies. The properties were then mapped to data relationships. After studying the steps, it analyzed the elements regarding the scope of the FAIR Principles and best practices of the W3C. As a result of the research, it established a set of semantic guidelines composed of elements and technologies that reflect the reality of laboratory research and the description of experiments with a plurality of attributes, precise and relevant, which may provide benefits to the scientific community with organized, standardized data and available for reuse. The proper application of these guidelines, in what concerns the structuring of data, collaborate so that the data are findable, accessible, interoperable and reusable.

Keywords: Scientific research data. Open notebook science. Laboratory notebooks. Linked Data. Semantic Web. E-Science. Semantic guidelines.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AACR2	-	Anglo-American Cataloguing Rules, segunda edição
ABEC	-	Associação Brasileira de Editores Científicos
ACD	-	*All Content – Delayed*
ACI	-	*All Content – Immediate*
API	-	*Application Programming Interface*
AVP	-	Ambiente Virtual de Pesquisa
BDTD	-	Biblioteca Digital Brasileira de Teses e Dissertações
BRAPCI	-	Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
CC	-	*Creative Commons*
CaFe	-	Comunidade Acadêmica Federada
CC0	-	Creative Commons CCZero
CC-BY	-	*Creative Commons Atribution*
CC-BY-NC	-	*Creative Commons License – No commercial*
CC-BY-NC-SA	-	*Creative Commons License – No Commercial / Share alike*
CC-BY-ND	-	*Creative Commons License – No Derived*
CC-BY-SA	-	*Creative Commons Atribution Share-Alike*
CDD	-	*Collaborative Drug Discovery*
CENSA	-	Collaborative Electronic Notebook Systems Association
CERN	-	Conseil Européen pour la Recherche Nucléaire
CSL	-	*CineStyle Color Lookup Format*
CSTB	-	*Computer Science and Telecommunications Board*
CSV	-	*Comma Sepated Values*
CVD	-	Ciclo de Vida dos Dados
DataONE	-	*Data Observation Network for Earth*
DCAT	-	*Data Catalog Vocabulary*
DCC	-	*Digital Curation Centre*
DCCCuration	-	*Curation Lifecycle Model The Digital Curation Centre*
DCMI	-	*Dublin Core Metadata Initiative*
DC Terms	-	Termos de metadados DCMI
DOAJ	-	*Directory of Open Access Journals*

DOI	-	*Digital Object Identifier*
DQV	-	*Data Quality Vocabulary*
DWBP	-	*Data on the Web Best Practices*
E-R	-	Entidade – Relacionamento
EmerRI	-	*Emerging Research Information*
FAIR	-	*Findable, Accessible, Interoperable, Re-usable*
FAPESP	-	Fundação de Amparo à Pesquisa do Estado de São Paulo
FOAF	-	*Friend of a Friend*
FORCE11	-	*Future of Research Communications and e-Scholarship*
FRAD	-	*Functional Requirements for Authority Data*
FRBR	-	*Functional Requirements for Bibliographic Records*
FRSAD	-	*Functional Requirementsfor Subject Authority Data*
FTP	-	*File-Transfer Protocols*
GBIF	-	*Global Biodiversity Information Facility*
GML	-	*Geography Markup Language*
HTML	-	*Hyper Text Markup Language*
HTTP	-	*Hyper Text Transfer Protocol*
IBICT	-	Instituto Brasileiro de Informação em Ciência e Tecnologia
IFLA	-	*International Federation of Library Associations and Institutions*
IFLA LRM	-	IFLA *Library Reference Model*
ISBD	-	*International Standard Bibliographic Description*
ISBN	-	*International Standard Book Number*
ISO	-	*International Organization for Standardization*
ISSN	-	*International Standard Serial Number*
JSON	-	*JavaScript Object Notation*
JSON-LD	-	*JavaScript Object Notation for Linked Data*
LCSH	-	*Library of Congress Subjetct Headings*
LISA	-	*Library and Information Science Abstracts*
LISTA	-	*Library, Information Science & Tecnology Abstracts*
LOD	-	*Linked Open Data*
LOV	-	*Linked Open Vocabularies*
MARC	-	*Machine Readable Cataloging*
MeSH	-	*Medical Subject Heading*

MP	-	Melhores Práticas para Dados na Web
NRC	-	*National Research Council*
NSB	-	*National Science Board*
NSF	-	*National Science Foundation*
OAI-PMH	-	*Open Archives Initiative Protocol for Metadata Harvesting*
ODbL	-	*Open Data Commons Open Database License*
ODC	-	*Open Data Commons*
ODC PDDL	-	*Open Data Commons Public Domanin Dedication and License*
ODC-BY	-	*Open Data Commons Attribution License*
ODS	-	*Open Document Spreadsheet*
OECD	-	*Organization for Economic Cooperation and Development*
OGP	-	*Open Graph Protocol*
OKF	-	*Open Knowledge Foundation*
ONS	-	*Open Notebook Science*
ONSchallenge	-	*Open Notebook Science Challenge*
ONSNetwork	-	*Open Notebook Network*
ORCID	-	*Open Research and Contributor ID*
OSM	-	*Open Source Malaria*
OWL	-	*Web Ontology Language*
PDDL	-	*Public Domain Dedication and Lincense*
PDF	-	*Portable Document Format*
PGD	-	Plano de Gestão dos Dados
PONS	-	*Partial Open Notebook Science*
RDA	-	*Resource Description and Access*
RDF	-	*Resource Description Framework*
RDFa	-	*Resource Description Framework in atributes*
RDFS	-	*RDF Schema*
re3data	-	*Registry of Research Data Repositories*
REST	-	*REpresentational State Transfer*
RSL	-	Revisão Sistemática de Literatura
SCI	-	*Selected Content – Immediate*
SCI-D	-	*Selected Content – Delayed*
SCI-I	-	*Selected Content – Immediate*

SDF	-	*Standard Database Format*
SGC	-	*Structural Genomics Consortium*
SKOS	-	*Simple Knowledge Organization System*
SOAP	-	*Simple Object Access Protocol*
SPARQL	-	*Protocol and RDF Query Language*
StART	-	*State of the Art through Systematic Review*
SVG	-	*Scalable Vector Graphics*
TIC	-	Tecnologias da Informação e Comunicação
Turtle	-	*Terse RDF Triple Language*
UNESP	-	Universidade Estadual Paulista Júlio Mesquita Filho
URI	-	*Uniform Resource Identifier*
VIAF	-	*Virtual International Authority File*
VRE	-	*Virtual Research Enviroments*
W3C	-	*World Wide Web Consortium*
XML	-	*EXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

O momento atual é marcado por uma avalanche de produção de dados, pela exploração de tecnologias de larga escala e pela colaboração entre diversos domínios do conhecimento. Esse fenômeno é denominado e-*Science* ou quarto paradigma da Ciência, o qual tem influenciado na forma de produção e comunicação da pesquisa cientifica, demandando, sobretudo, a abertura de seus dados e não apenas de seus resultados. Acredita-se que o uso inteligente dos elementos que assinalam a *e-Science* pode aumentar a velocidade de circulação da informação, possibilitando a colaboração em rede e acelerando novas descobertas. Porém, a capacidade de acompanhar e se beneficiar das propostas da *e-Science* depende do desenvolvimento de ações para capturar, gerenciar, analisar e disponibilizar informação à comunidade científica.

A publicação de dados de pesquisa científica encontra aportes teóricos e práticos na linha de pesquisa que integra a Informação e a Tecnologia, a qual é objeto de estudo da Ciência da Informação, e busca refletir sobre as transformações sociais e tecnológicas para a construção do conhecimento em torno de novas formas de acesso e uso da informação (UNESP, 2019).

Para Santarem Segundo (2010), a Ciência da Informação tem participado efetivamente dessa transformação, alavancada pelo uso intensivo de recursos tecnológicos, os quais facilitam o processo de disseminação da informação, incluindo o conhecimento cientifico, que deixou de estar disponível apenas em formatos impressos e passou a utilizar a estrutura tecnológica para organizar e disponibilizar a informação registrada em plataformas digitais.

A Ciência da Informação, assim denominada a partir do final da década de 1960, preocupa-se com a informação científica desde a sua origem até o seu uso. Para Borko (1968) a Ciência da Informação é a disciplina que investiga as propriedades e o comportamento da informação, as forças que governam o fluxo informacional e os meios para processá-la a visando à otimização do acesso e uso. Está relacionada com um corpo de conhecimento que abrange entre outras ações a origem, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação. Ainda segundo Borko (1968), esse corpo de conhecimento inclui a investigação de três tipos de fenômenos: a representação da informação em sistemas naturais e artificiais; o relacionamento com o uso de códigos para transmissão eficiente da mensagem e o estudo dos meios e técnicas eficientes da mensagem.

Os estudos na área da Ciência da Informação foram se ampliando ao longo do tempo e se beneficiando de outras áreas. Nessa direção, Foskett (1980) refere-se à Ciência da Informação com foco na interdisciplinaridade e na transferência do conhecimento organizado ao mencionar o surgimento de uma fertilização cruzada de ideias que incluem a velha arte da biblioteconomia, a nova arte da computação, as artes dos novos meios de comunicação e aquelas ciências como psicologia e linguística, que, em suas formas modernas, têm a ver com todos os problemas da comunicação – a transferência do conhecimento organizado.

Além da interdisciplinaridade, Saracevic (1996) destaca a ligação inexorável com a tecnologia da informação e a participação ativa na evolução da sociedade da informação como características fundamentais da Ciência da Informação.

A partir das palavras de Borko (1968), Foskett (1980) e Saracevic (1996) vislumbra-se nos fundamentos da Ciência da Informação o caminho para a publicação de dados de pesquisa científica por meio da implementação das ações do fluxo informacional associadas ao uso intensivo de tecnologias da informação e comunicação, expandidas a partir do final da década de 1990.

A publicação de dados de pesquisa, em formato aberto, exige tecnologias que permitem a colaboração aberta e simultânea. Em se tratando de cadernos abertos de pesquisa ou *Open Notebook Science*, objeto de estudo desta tese, busca-se tecnologias, vocabulários padronizados e amplamente reconhecidos para possibilitar a reutilização e apoiar descobertas.

Nota-se que as características apresentadas, pelos autores anteriormente mencionados, permanecem essencialmente nos projetos da Ciência da Informação, enquanto que as maneiras de executar as suas atividades vem se modernizando conforme demandas do momento. Para Araújo (2018), a Ciência da Informação vem evoluindo por meio das suas subáreas, como por exemplo, a Biblioteconomia e a Computação, bem como pelas tentativas de caracterização de campo como ciência interdisciplinar, social e pós-moderna.

Nesse contexto, observa-se uma abordagem atual da área apresentada por Santos e Sant'Ana (2013, p. 200) ao mencionarem que

> A Ciência da Informação refere-se à atividade direcionada à pesquisa de princípios e métodos que são partes da análise, do projeto e da evolução dos sistemas de informação. Nesses sistemas, os elementos constituintes são o ambiente, as pessoas, os recursos informacionais, as tecnologias e os procedimentos. Eles sustentam a capacidade para a busca de soluções e tomada de decisões como parte da vida diária, envolvendo a manipulação de dados, o acesso à informação e a apropriação do conhecimento.

Os sistemas de informação e seus elementos, apresentados por Santos e Sant'Ana (2013), estão presentes no desenvolvimento da proposta desta pesquisa, *diretrizes semânticas*

para publicação de dados de pesquisa científica de cadernos de laboratório, a qual requer o conhecimento das necessidades de consumo da sociedade, no levantamento de conceitos e tecnologias para tipos de conjuntos de dados e na infraestrutura disponível. Esse processo envolve pessoas, pesquisas de princípios, tecnologias, métodos e procedimentos. No que se refere às pessoas, este estudo descreve os dados a partir do relacionamento entre dados de pesquisa e às necessidades dos usuários, com base em modelo entidade-relacionamento; as etapas das diretrizes propostas são construídas a partir de princípios e tecnologias para publicação de dados abertos de pesquisa, de modo a contribuir para que os mesmos fiquem facilmente recuperáveis e reutilizáveis por pessoas com interesses semelhantes.

O produto desta tese é essencialmente conceitual, para ser adotado em plataformas digitais, com vista à estruturação e publicação de dados de pesquisa científica anotados em cadernos de pesquisa à luz das dimensões da *e-Science*, de maneira a tornar as informações compreensíveis por usuários humanos e máquinas, e favorecer a colaboração científica em rede entre pesquisadores de diferentes domínios do conhecimento. Esse conjunto de diretrizes semânticas pretende apoiar as novas práticas científicas, contribuir para a modernização da Ciência da Informação e consequentemente maximizar os avanços da ciência.

1.1 PROBLEMA DA PESQUISA

A questão problema a ser investigada consiste em: como publicar dados de pesquisa científica anotados em cadernos de pesquisa, em formato aberto e semântico, que atenda as demandas da *e-Science*?

Esse problema de pesquisa pode ser desmembrado nos seguintes questionamentos:

1 - Quais elementos conceituais e práticos presentes nas dimensões da *e-Science* precisam ser observados para publicação de dados de pesquisa científica?

2 – Quais diretrizes e tecnologias semânticas podem ser adotadas para publicar conjuntos de dados de cadernos de pesquisa em plataformas de acesso aberto?

3 – Quais etapas e elementos são necessários para compor um conjunto de diretrizes para estruturação e publicação de dados de pesquisa científica em formato aberto e semântico?

Entende-se que a provável resposta ao problema de pesquisa tem como base os elementos apresentados na tese e na hipótese a seguir.

1.2 TESE E HIPÓTESE

A tese formulada nesta pesquisa sugere que a publicação de dados de pesquisa científica, quando realizada de forma estruturada, aberta e semântica, contribua para melhorias do processo de publicação científica. Em se tratando de dados gerados em laboratórios de pesquisa, a publicação dos dados experimentais evita a duplicidade de esforços, amplia a transparência na sua divulgação e possibilita a colaboração em rede.

A partir dessa proposição, a hipótese construída para esta pesquisa é que um conjunto de diretrizes, elaborado a partir de 1) elementos conceituais e práticos da *e-Science*, 2) princípios do *Linked Data* e tecnologias da Web Semântica, 3) uso do modelo conceitual IFLA LRM para delimitar a estruturação de descrição, 4) princípios FAIR (encontráveis, acessíveis, interoperáveis e reutilizáveis) e adoção de boas práticas para publicação de dados na Web, poderá contribuir para a publicação de cadernos abertos de pesquisa propiciando a reutilização de dados, a ligação de recursos ricamente descritos na Web e poderá favorecer a circulação da informação e novas descobertas.

Acredita-se que essa hipótese poderá ser verificada por meio dos objetivos a seguir.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Propor um conjunto de diretrizes semânticas para estruturação e publicação de dados abertos de cadernos de pesquisa, visando às melhorias na recuperação e compartilhamento de dados em plataformas de acesso aberto.

1.3.2 Objetivos Específicos

1 - Analisar, com base na literatura, os elementos conceituais e práticos presentes nas dimensões da *e-Science*.

2 - Apresentar as características e especificidades dos dados de pesquisa científica anotados em cadernos de pesquisa.

3 – Analisar, a partir de iniciativas existentes, as práticas favoráveis e as desfavoráveis na publicação de dados de pesquisa científica anotados em cadernos abertos de pesquisa.

4 - Identificar conceitos e tecnologias da Web Semântica e *Linked Data* para publicar conjuntos de dados de cadernos de pesquisa em plataformas de acesso aberto.

5 - Identificar etapas e elementos para compor o conjunto de diretrizes para estruturação e publicação de dados de pesquisa científica de cadernos de laboratório em formato aberto, semântico e que atendam aos princípios FAIR.

1.4 JUSTIFICATIVA, RELEVÂNCIA E MOTIVAÇÃO DA PESQUISA

O ecossistema da pesquisa cientifica vem passando por diversas e rápidas transformações que interferem em como a informação é gerada e compartilhada. Essa transformação está ocorrendo, em âmbito mundial, por uma gama de fatores, desde os avanços na tecnologia até pressões de financiamento e cultura colaborativa entre pesquisadores. Esse conjunto de fatores demanda estudos e planejamento para acompanhar e interferir em tais mudanças e contribuir para as novas formas de se fazer ciência.

Essa nova forma de fazer ciência contempla a divulgação de dados primários, preferencialmente, em tempo real, na medida em que são gerados, e não apenas os casos de sucesso ou de resultados consolidados. A tendência é permitir a colaboração simultânea, à distância e, cada vez mais, de modo visível e aberto à ampla contribuição. Essa tendência é válida, pois segundo Curty (2015, p. 5), "supõe-se que dados podem ser úteis para outras pessoas – dentro e fora de domínios disciplinares – e, portanto, melhora as chances de novos resultados e conhecimentos científicos decorrentes dos mesmos dados já disponíveis".

Os cadernos de pesquisa ou *notebook science* são instrumentos de anotações de dados experimentais gerados em laboratórios para fundamentar as publicações científicas, que são divulgadas normalmente em suas configurações finais, seja no formato de teses ou dissertações, publicações de livros, artigos de periódicos e comunicações de congressos. Para Schnell (2015) o caderno de laboratório registra as hipóteses, experimentos e análises iniciais ou interpretações dos experimentos; e serve como um registro legal de propriedades das ideias e resultados obtidos por um cientista. Apesar da importância atribuída aos dados registrados em cadernos de pesquisa, tais dados, em sua maioria, não são divulgados juntamente às publicações finais, ficando inacessíveis aos demais pesquisadores e à sociedade.

Nesse sentido, vale enfatizar que, em tempos de *e-Science*, não basta anotar os experimentos em cadernos de pesquisa, comumente chamados de cadernos de laboratório, é preciso disseminar, em formato aberto, o conhecimento produzido por instituições de pesquisa. Esse movimento traz benefícios à comunidade cientifica global, uma vez que valida os resultados das pesquisas e aumenta a transparência, principalmente, quando se tratam de pesquisas financiadas por instituições públicas. Além disso, reforça-se a ideia de que os dados de pesquisa são parte integrante do registro acadêmico e devem estar disponíveis para reuso.

Sabe-se que o desenvolvimento de infraestruturas sob o paradigma emergente da *e-Science* permite que dados sejam compartilhados e acessados de forma aberta. Contudo, não é o suficiente desenvolver ambientes digitais para publicação de dados que sejam apenas

plataformas para estocá-los, faz-se necessária a aplicação de metodologias construídas a partir de conceitos e tecnologias da Web Semântica e *Linked Data*, as quais permitam a interpretação por pessoas e máquinas, além de possibilitar ligações entre recursos existentes e apoiar o enriquecimento de dados por meio do estabelecimento de novas relações entre recursos.

Sendo assim, o desenvolvimento de um conjunto de diretrizes semânticas para a publicação desses dados de pesquisa científica anotados em cadernos de pesquisa se justifica para criar ambientes abertos e expansíveis para facilitar a pesquisa, de forma que outras pessoas possam contribuir e acrescentar esforços em todos os seus estágios.

Nesse sentido, para desenvolver a proposta em questão é preciso prever os elementos e etapas conceituais para futuras implementações em plataformas digitais. Para isso, está prevista a identificação do ecossistema dos cadernos de pesquisa, a modelagem dos dados e o mapeamento de metadados e vocabulários para descrição dos referidos cadernos.

Para a modelagem dos dados de pesquisa será adotado o modelo conceitual IFLA LRM, por ser do tipo entidade-relacionamento, com foco nas tarefas dos usuários (encontrar, identificar, selecionar, obter e explorar), contemplando os aspectos do universo bibliográfico e oferecendo uma estrutura de mapeamento de atributos e sugestão de valores para os metadados. Nesse contexto, vale ressaltar que o AACR2, em 1978, definiu em seu glossário que recurso bibliográfico é a expressão ou manifestação de uma obra ou um item que constitui a base de uma descrição. Seguindo esse raciocínio, tem-se que os dados de pesquisa podem ser uma forma de expressão da obra. Cabe salientar que nem todos os dados de pesquisa de cadernos de laboratório podem ser considerados dados bibliográficos. No entanto, o fato da proposta desse modelo conceitual ser centrado nas tarefas do usuário, considera-se relevante a sua adoção para apoiar na definição de metadados e relacionamentos entre os elementos presentes nesses cadernos, o que facilita o processo de mapeamento de propriedades.

É uma proposta considerada promissora no modo de produzir e comunicar a ciência, uma vez que foi constatada, por meio de levantamento de fontes primárias, a ausência de pesquisas que englobam as temáticas Caderno de Pesquisa, Princípios FAIR, Web Semântica e *Linked Open Data,* com a apresentação de dados estrutrados para efeito de publicação em plataformas de acesso aberto.

Para o estudo e análise de trabalhos correlatos, realizou-se o processo de busca e recuperação de fontes primárias, baseado na metodologia de revisão sistemática de literatura e indicações de especialistas no assunto, conforme apresentado no item 2.4.

Na literatura brasileira, há estudos de destaque e com grande relevância em termos de abrangência para a fundamentação teórica e construção dos capítulos desta tese. Entre esses trabalhos, contemplou-se seis teses que possuem como tema central o estudo dos dados de pesquisa científica e uma dissertação, que discute a *e-Science* e as práticas de pesquisa cientifica. Os autores relacionados são Appel (2014), Sales (2014), Curty (2015), Clinio (2016), Oliveira (2016), Costa (2017) e Semeler (2017).

Em 2014, em um estudo de caso no *Conseil Européen pour la Recherche Nucléaire* (CERN), Appel analisou a relação entre novas práticas de produção colaborativa de conhecimento científico e o desenvolvimento e uso de plataformas tecnológicas de amparo à pesquisa colaborativa, levando em consideração as diferentes visões, perspectivas e interesses dos atores atuantes nessas práticas. Analisou, ainda, as definições quanto aos direitos de acesso e uso dos dados de pesquisa em tais práticas.

Ainda em 2014, Sales propôs a integração semântica de publicações científicas e dados de pesquisa para a área de Ciências Nucleares. Como resultados obteve uma proposta de diretrizes para uma política nacional de curadoria digital e um modelo de publicação científica para área de Ciências Nucleares. Nessa perspectiva, Sales (2014) adotou princípios semânticos para a construção do modelo, mas não aprofundou nas especificidades dos cadernos de pesquisa.

Em 2015, Curty identificou os fatores que influenciam a reutilização de dados de pesquisa nas Ciências Sociais, por meio de um estudo realizado na Universidade de Syracuse, em NY, Estados Unidos. Os resultados forneceram um entendimento sobre a reutilização desses dados no contexto da ciência aberta e forneceram elementos que contribuem para as decisões dos cientistas sociais para reutilizar dados coletados por outros. Essa pesquisa, contudo, não aborda infraestruturas para a publicação de dados em questão.

Em 2016, Clinio pesquisou sobre os novos cadernos de laboratório e analisou novas culturas epistêmicas, as quais estão sendo engendradas por dois modos de ciência emergentes – a Ciência Aberta e a Ciência Comum. Apesar de inúmeras diferenças, essas ciências convergem na crítica à noção de fato científico e na estratégia de transformar o caderno de laboratório em sua principal tecnologia literária. Nessa pesquisa, Clinio (2016) menciona a preocupação de Jean-Claude Bradley em adotar princípios da Web Semântica em suas iniciativas para dialogar com outros grupos que lidam com a questão da representação da informação em Química. Em uma busca mais detalhada sobre as iniciativas de Bradley, identificou-se o projeto *UsefulChem*, hospedado em um site gratuito de *Wikispaces*, porém esse serviço foi descontinuado em janeiro de 2019.

Ainda em 2016, a pesquisa de Oliveira busca elaborar padrões que promovam a recompensa autoral na *e-Science*. O estudo desse autor está inserido no contexto da ciência aberta, *e-Science*, direitos de propriedade intelectual, dados de pesquisa científica e autoralidade colaborativa. Oliveira (2016) concluiu que a autoralidade no contexto da *e-Science* é colaborativa, a qual é garantida mediante recompensas por meio de atribuição, citação e responsabilização. Oliveira (2016) explica que a responsabilidade da atribuição e citação no cenário contemporâneo é atribuída a cada colaborador de acordo com sua participação na pesquisa. Assim, a tese se confirmou e está representada pelo modelo conceitual de autoralidade colaborativa na *e-Science*.

Em 2017, Semeler investigou como as relações entre Ciência da Informação, *e-Science* e *Data Science* influenciam a Biblioteconomia de Dados. A pesquisa perpassa pela apresentação de repositório de dados de pesquisa e busca fundamentação teórica nos temas sobre informação, tecnologia e aspectos gerais da Ciência da Informação; conceitos de dados, incluindo dados de pesquisa, as diferentes disciplinas e biblioteconomia de dados; *data Science*; e, tecnologias e instrumentos para manipulação de dados. A pesquisa finaliza com uma discussão a respeito do Diagrama de Venn da Biblioteconomia de Dados, o qual representa as bases da Ciência da Informação, *e-Science* e *Data Science* como fundamentos para a Biblioteconomia de Dados.

Por último, em 2017, Costa analisou as diretrizes para uma política de gestão de dados científicos no Brasil, concluindo que uma política dessa natureza gestão precisa abordar aspectos, tais como: regras de compartilhamento e reuso dos dados, prazo de carência para algumas categorias de dados, prazo de armazenamento para algumas classes de dados, padrões de metadados e a interoperabilidade destes.

Além dos estudos realizados em teses e dissertações, outras pesquisas foram publicadas em livros e artigos científicos sobre os temas: Web Semântica e *Linked Data*, *e-Science*, dados de pesquisa científica e cadernos de laboratório. No entanto, reforça-se a informação de que tais pesquisas não referem-se à publicação de cadernos de pesquisa a partir de tecnologias da Web Semântica e conceitos do *Linked Data*. Apesar disso, muitas pesquisas motivaram e colaboraram para o desenvolvimento da proposta desta tese.

Os textos de Jean-Claude Bradley, principal autor de *Open Notebook Science*, foram as principais fontes de informação para a escolha do tema, bem como os diversos textos de Clinio e Albagli que colaboraram com a concepção da proposta desta tese, no que se trata ao objeto da pesquisa, ou seja, a motivação para estudar sobre os cadernos abertos de pesquisa.

Diversos textos que abarcam o estudo sobre dados de pesquisa referem-se a tipologias, princípios e diretrizes, gestão e curadoria de dados, compartilhamento de dados, os quais mencionam a importância da publicação de dados estruturados, porém não detalham sobre a publicação de cadernos abertos de pesquisa. Dentre eles incluem: Organization for Economic Co-Operation and Development (2007), Borgman (2010), Tenopir (2011; 2015), FORCE11 (2014), Appel (2014), Sayão e Sales (2015), Sant'Ana (2016), Ferreira (2018) e Silva (2019).

Sobre a infraestrutura da *e-Science*, tecnologias da Web Semântica e *Linked Data* relacionada à publicação de dados destacam-se Gray (2009), Hey, Tansley e Tolle (2009), Santarem Segundo (2014; 2015; 2018), Consórcio W3C, Appel (2014), Laufer (2015), Schapira (2019), Isotani e Bittencourt (2019), Rautenberg, Souza, Dall'Agnol e Michelon (2018), entre outros já citados no decorrer desta pesquisa.

As pesquisas mencionadas que abordam o uso de conceitos e tecnologias da Web Semântica, por meio do *Linked Data*, para publicação específica de conjuntos de dados anotados em cadernos de pesquisa, não apresentam a estrutura como vocabulários, metadados, esquemas de metadados, incluindo classes e propriedades apropriadas para a publicação de cadernos de pesquisa.

A partir das premissas analisadas e o levantamento bibliográfico realizado na revisão sistemática de literatura, justifica-se a elaboração de um conjunto de diretrizes semânticas para estruturação que visa a explicitação dos os elementos apropriados para a publicação de dados abertos de cadernos de pesquisa e para facilitar a ação de encontrar, interoperar, acessar, compreender e reutilizar dados de pesquisa gerados em laboratórios.

Vale destacar que as principais motivações para realização desta pesquisa consistem em ter pleno conhecimento da importância da divulgação de dados de pesquisa científica anotados em cadernos de pesquisa e saber que ainda não possui um conjunto diretrizes semânticas para apoiar na estruturação e publicação desses dados em uma plataforma de acesso aberto. Motiva também o desejo de colaborar para a produção de novos conhecimentos que integram os elementos Dados de Pesquisa Cientifica, Cadernos de Pesquisa, Web Semântica, conceitos do *Linked Data* e Princípios FAIR.

Por fim, espera-se que as diretrizes propostas possam apoiar os cientistas com dados confiáveis e de qualidade, cooperando, assim, para a gestão de dados de pesquisa científica que vem sendo exigida por agências de fomento e instituições de ensino superior. Espera-se, ainda, que essa iniciativa possa colaborar para o projeto Ciência Aberta no Brasil e para a expansão da literatura nas áreas da Ciência da Informação e da Computação.

2 METODOLOGIA E ESTRUTURA DA TESE

De acordo com Demo (2012, p. 11), a metodologia significa "na origem do termo, [o] estudo dos caminhos, dos instrumentos usados para se fazer ciência". Para o autor, é uma disciplina instrumental que visa, também, a indagar os limites da ciência, seja para conhecer a realidade do objeto, seja para intervir na realidade do objeto.

Essa definição representa a proposta da tese, pois ao construir um conjunto de diretrizes semânticas para estruturar e publicar cadernos de pesquisa, o estudo busca sistematizar caminhos para se colaborar com a ciência e acompanhar a demanda dos novos tempos, carregada de avanços tecnológicos, necessidades rápidas de produção e descobertas do conhecimento científico.

O estudo está sendo norteado pela questão problema da pesquisa e pautado em abordagens teóricas e práticas que fundamentam a elaboração das diretrizes citadas, as quais buscam, ao serem implementadas melhorias na qualidade da recuperação, acesso, uso e compartilhamento de dados em plataformas de acesso aberto.

Para o desenvolvimento desta tese, a metodologia foi subdividida em duas etapas:

1 – Procedimentos metodológicos tradicionais – para a delimitação do universo e amostragem da pesquisa, classificação dos tipos de pesquisa e coleta dos dados sobre a tipologia e características dos dados dos cadernos selecionados para estudo;

2 – Revisão Sistemática de Literatura – por meio do processo de busca e recuperação de fontes primárias, com vista a identificar trabalhos relacionados ao estado da arte dos cadernos abertos de pesquisa, no que se refere a sua publicação e formas de acesso e uso.

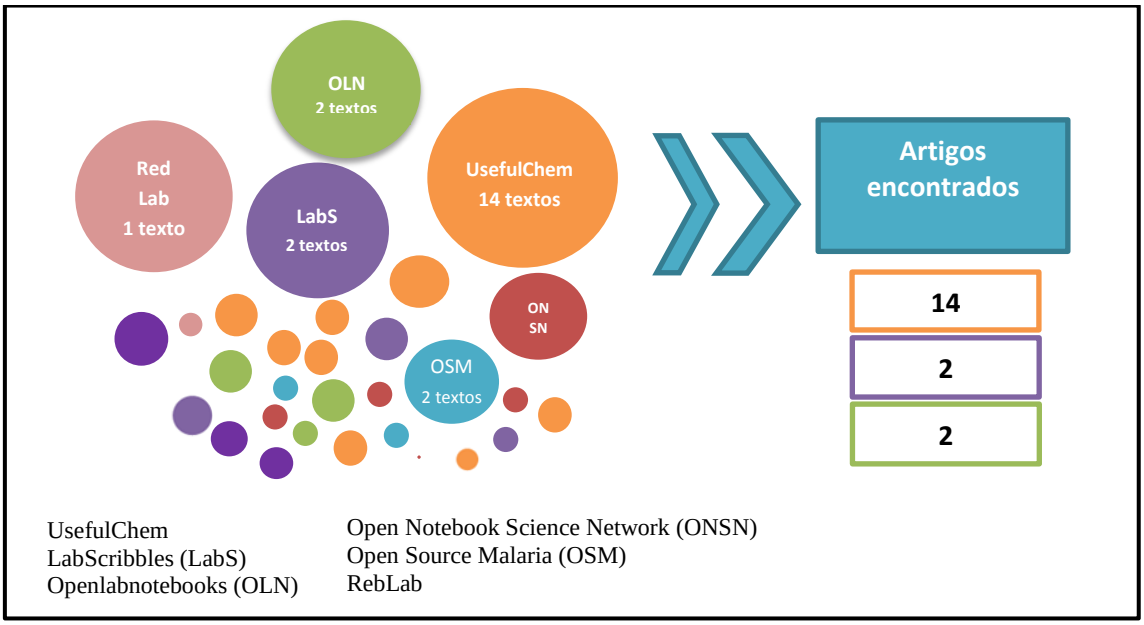
2.1 DELIMITAÇÃO DO UNIVERSO E AMOSTRAGEM DA PESQUISA

O universo em estudo trata-se dos dados, sem análise e inferências do autor, gerados no transcurso de uma pesquisa científica. A amostra selecionada dentro desse universo refere-se aos dados da pesquisa científica anotados em cadernos abertos de pesquisa.

Para caracterizar a tipologia dos dados que os cadernos de pesquisa publicam, optou-se por analisar os cadernos que estivessem ativos no momento da realização desta tese, bem como o de maior destaque na literatura selecionada, considerando os critérios predefinidos na revisão sistemática de literatura. A partir desses critérios, selecionou-se os cadernos *UsefulChem*, *LabScribbles* e *Openlabnotebooks* para análise das características e tipologias de

dados de pesquisa científica. Os demais cadernos citados na literatura, embora que apenas exemplificados, foram apresentados no item 5.3.3.

Figura 01 - Amostragem da pesquisa - cadernos abertos de pesquisa



Fonte: Elaborado pela autora (2020).

O *UsefulChem* foi descrito ou exemplificado em 14 estudos primários identificados na revisão sistemática de literatura. Apesar de o projeto *UsefulChem* ter sido descontinuado em fevereiro de 2019, é o caderno de maior destaque na literatura sobre projetos *Open Notebook Science*, por ser a primeira iniciativa de caderno aberto de pesquisa, registrado na literatura (BIRD; WILLOUGHBY; FREY (2013); BIRD; COLES; FREY (2015), MILSTED (2015), CLINIO (2016)). No entanto, os seus *datasets* foram identificados no repositório re3data.org, o que possibilitou a análise das características e tipologias dos dados.

A escolha pelos cadernos *LabScribbles* e *Openlabnotebooks* deu-se por serem citados na literatura e por estarem ativos em um repositório de dados para armazenamento e divulgação, o que proporcionou conhecer a estrutura dos dados publicados.

Os projetos de cadernos abertos de pesquisa *Open Source Malaria*, *Open Notebook Science Network* (ONSN) e RebLab são exemplificados na literatura e, por isso, foram listados no item 5.3.3, porém não foram objetos de análise desta tese.

Os cadernos identificados na literatura são em sua maioria da área química. O fato de Jean-Claude Bradley, precursor da abertura de caderno de pesquisa, ser químico, a principal

literatura sobre o tema está atrelada a proposta e exemplos dessa área. Neste sentido, os mapeamentos realizados no decorrer desta tese também se aproximam da área química.

2.2 CLASSIFICAÇÃO DA PESQUISA CIENTÍFICA

Para Gil (2016) uma forma de organizar os fatos e melhorar o entendimento do conhecimento humano é classificar a pesquisa. Sendo assim, esta pesquisa foi classificada segundo a finalidade aplicada, a natureza qualitativa, o método bibliográfico e os objetivos descritivos, exploratórios, levantamento e documental.

2.2.1 Pesquisa Aplicada

Para Gil (2016) a pesquisa, segundo a sua finalidade aplicada, se destina à aquisição de conhecimentos com vista à aplicação numa determinada situação. As diretrizes - enquanto produto final desta tese - possui a finalidade de serem aplicadas em plataformas digitais para gerenciamento e publicação de dados científicos de cadernos de pesquisa.

2.2.2 Pesquisa Qualitativa

Segundo Creswell (2010), a pesquisa de natureza qualitativa refere-se à investigação baseada em análise e interpretação de dados textuais, imagens, relatos e opiniões que descrevem o objeto a ser analisado. A pesquisa desta tese é de natureza qualitativa por se tratar da análise das características inerentes ao objeto em estudo (dados científicos de cadernos de pesquisa) e de um conjunto de tecnologias, padrões e boas práticas para constituir um conjunto de diretrizes semânticas adequadas para publicação de tais dados de maneira a serem encontráveis, acessíveis, interoperáveis e reutilizáveis.

2.2.3 Pesquisa Descritiva

A pesquisa descritiva é adotada em estudos que pretendem realizar o levantamento das características dos componentes do fato, fenômeno ou processo. A metodologia descritiva está aplicada na construção dos capítulos referentes à revisão de literatura e à construção das diretrizes, a qual descreve as características de todos os elementos que compõem a tese.

2.2.4 Pesquisa Exploratória

Para Gil (2016, p. 27) as pesquisas exploratórias têm como objetivo “proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explicito ou a construir

hipóteses". Serra (2019) destaca as características da pesquisa exploratória ao mencionar a análise feita por Raupp e Bueren (2006, p.80)

> [...] a pesquisa exploratória normalmente ocorre quando há pouco conhecimento sobre a temática a ser abordada. Por meio de estudo exploratório, busca-se conhecer com maior profundidade o assunto, de modo a torná-lo mais claro ou construir questões importantes para a condução da pesquisa.

Diante das características apresentadas, adota-se a pesquisa exploratória para conhecer com maior profundidade os temas que contemplam o segundo, terceiro e quarto objetivo específico deste estudo, sendo que o segundo objetivo busca apresentar as especificidades dos dados de pesquisa anotados em cadernos de laboratório, o terceiro refere-se ao estudo das práticas favoráveis e desfavoráveis de iniciativas existentes e o quarto objetivo busca identificar conceitos e tecnologias da Web Semântica e *Linked Data* para publicar conjuntos de dados científicos de cadernos de pesquisa em plataformas de acesso aberto.

2.2.5 Pesquisa Bibliográfica

A pesquisa bibliográfica refere-se ao procedimento que visa a identificar, consultar e analisar documentos publicados por autores proeminentes da área para fins de embasamento teórico da pesquisa. Neste estudo, a pesquisa bibliográfica foi construída a partir de levantamentos de publicações que abarcam os temas *e-Science*, Dados de Pesquisa Científica, Cadernos abertos de pesquisa, Web Semântica e *Linked Data*. As pesquisas foram realizadas em bases de dados nacionais e internacionais - BDTD, BRAPCI, DOAJ, LISA, LISTA, Scopus, Web of Science e Google Scholar -, nos idiomas português, inglês e espanhol, considerando o período cronológico de buscas estabelecidas conforme o assunto, os quais são informados na etapa seguinte da pesquisa de levantamento. A tipologia de documentos para a coleta de bibliografia correspondeu a livros, capítulos de livros, artigos científicos, teses, dissertações, relatórios técnicos e sites oficiais.

2.2.6 Pesquisa de Levantamento

A metodologia de levantamento foi adotada nesta pesquisa com os seguintes propósitos:

1 – Levantamento e análise de material bibliográfico sobre *e-Science* e os principais elementos de suas dimensões – consistiu em um levantamento realizado nas bases já indicadas, no período de 2000 a 2018, período de realização do capítulo.

2 – Levantamento e análise de material bibliográfico sobre dados de pesquisa científica, suas tipologias, modelos de ciclo de vidas dos dados e diretrizes com orientações para estruturação e publicação de tais dados – o levantamento foi realizado nas mesmas bases de dados indicadas e, principalmente, por recomendação de especialistas, estabeleceu ao final da década de 90 e acompanhou a evolução do tema até 2019.

3 – Levantamento, identificação e análise de material bibliográfico sobre cadernos abertos de pesquisa – por se tratar do objeto de pesquisa da tese, o levantamento foi subsidiado pela revisão sistemática de literatura, que buscou identificar o estado da arte dos cadernos abertos de pesquisa em relação a sua estrutura e publicação.

4 – Levantamento e análise do material bibliográfico sobre Web Semântica e *Linked Data* – consistiu em um levantamento das bases de dados indicadas no item sobre pesquisa bibliográfica e seleção, com o apoio de especialistas e professor orientador, durante o período de busca que correspondeu de 2001até a evolução do tema, em 2020.

5 – Levantamento de dados – refere-se à etapa prática da pesquisa que buscou compreender e caracterizar os tipos de dados de pesquisa. Os conjuntos de dados dos cadernos selecionados foram encontrados e verificados nas plataformas re3data, Zenodo e *Blogger*.

2.2.7 Pesquisa Documental

Vale-se de documentos que não receberam ainda um tratamento analítico, são os documentos considerados de primeira mão. Sendo assim, adotou-se a pesquisa documental na consulta realizada aos dados de cadernos de pesquisa, por meio das plataformas Zenodo e *Blogger*.

2.3 COLETA DE DADOS

Para compreender a tipologia e características dos dados de pesquisa científica dos cadernos de laboratório, fez-se uma consulta nas seguintes plataformas:

1 – Caderno *UsefulChem* (plataforma *Blogger* e *re3data*) – esse caderno dedicou-se a desenvolver estudos contra a malária. Por se tratar de um *blog* cada postagem contém um composto molecular seguido de informações. Não possui filtros de busca por tipos de documentos, contudo foi possível observar a publicação de vídeos, aulas (ppt), palestras (ppt), entrevistas, capítulos de livros (pdf) e artigos científicos (pdf) de autoria de Jean-Claude Bradley.

2 - Caderno *LabScribbles* (repositório de dados Zenodo) – esse caderno dedica-se a pesquisas para tratamento da doença de *Huntington*. Para esse caderno, os filtros disponíveis para consulta são: tipo de dados, tipo de arquivo e termo de busca. O termo adotado para consulta foi: *Huntington*. Não se estabeleceu datas, pois se pretendeu recuperar todos os tipos de documentos e formatos adotados para o armazenamento dos dados de pesquisa. As informações recuperadas foram:

Quadro 01 - Tipologia de dados e arquivos do Caderno *LabScribbles*

Tipo de Arquivo		**Tipos de Dados**			
Extensão do arquivo		**Conj. de dados (*dataset*)**	**Publicação** Artigo, Relatório, *Preprint*, *paper* de conferências, nota técnica e documento de trabalho	**Apresentação**	**Pôster**
docx	Word (versão 2007-2010)	118	6	0	0
pdf	Formato portátil de documento	43	1	3	2
xlsx	Excel (versão 2010)	13	0	0	0
ppt	Power Point	4	1	1	0
raw	Formato de imagem	3	0	0	0
xls	Excel (versão 2003-2007)	125	0	0	0
doc	Word (versão 97-2003)	1	0	0	0
sf3	Arquivo de som	125	0	0	0

Fonte: Repositório Zenodo (2020).

3 - Caderno *Openlabnotebooks* (repositório de dados Zenodo) – esse caderno também desenvolve pesquisas para tratamento da doença de *Huntington*. Os termos adotados para consulta foram: tipo de dados, tipo de arquivo e termo de busca. Os critérios de coleta dos dados foram os mesmos, portanto, não se estabeleceu datas, pois se pretendeu recuperar todos os tipos de dados e formatos adotados para o armazenamento dos dados de pesquisa. As informações recuperadas foram:

Quadro 02 - Tipologia de dados e arquivos do Caderno *Openlabnotebooks*

Tipo de Arquivo		**Tipos de Dados**			
Extensão do arquivo		**Conj. de dados (*dataset*)**	**Publicação** Artigo, Relatório, *Preprint*, *paper* de conferências	**Apresentação**	**Pôster**
docx	Word (versão 2007-2010)	67	3	0	0
pdf	Formato portátil de documento	38	0	2	1
xlsx	Excel (versão 2010)	8	2	0	0
ppt	Power Point	0	2	0	0
mtz	Compactar arquivos	68	0	0	0
xls	Excel (versão 2003-2007)	20	0	0	0
zip	Compactar arquivos	0	1	0	0

Fonte: Repositório Zenodo (2020).

As definições dos tipos de dados foram apresentadas na seção 5.4. Tais dados foram analisados e explorados no decorrer desta pesquisa.

2.4 REVISÃO SISTEMÁTICA DE LITERATURA

Segundo Kitchenham e Charters (2007), a revisão sistemática de literatura (RSL) é um método científico que visa a identificar, avaliar e sintetizar a pesquisa disponível em dado momento, sobre um tema específico, de forma objetiva e reproduzível. Para Galvão e Pereira (2014), uma das características da RSL é a definição de critérios de busca padronizada e o registro do que se faz para possibilitar a reprodução futura do estudo.

De acordo com Popay, Rogers e Willians (1998), o processo da RSL deve pautar critérios sistemáticos que garantam rigor metodológico. Para os autores, a RSL deve seguir um protocolo sistemático que indicará o que foi feito antes para sustentar a atividade proposta e deve apresentar as seguintes etapas: planejamento, condução e resultados, as quais serão descritas nas seções seguintes.

Para realizar as etapas da revisão sistemática desta tese, baseou-se nas propostas de Popay, Rogers e Willians (1998), Kitchenham e Charters (2007) e Galvão e Pereira (2014).

2.4.1 Planejamento da Revisão Sistemática de Literatura

O planejamento da RSL é a etapa que se define um protocolo formalizado com as atividades a serem seguidas na execução da revisão proposta. Os elementos que compõem a etapa do planejamento são: objetivo da RSL, definição da questão da pesquisa da RSL, critérios de seleção das bases de dados, de estratégias de busca, de seleção dos estudos primários (Inclusão e Exclusão), de seleção de estudos iniciais e critérios para avaliação da qualidade. Esses critérios serão detalhados a seguir:

a) Objetivo da revisão sistemática

Identificar trabalhos relacionados ao estado da arte dos cadernos abertos de pesquisa em relação a sua estrutura e publicação, bem como analisar tecnologias e boas práticas para publicar conjuntos de dados de pesquisa científica anotados em cadernos de pesquisa.

b) Questões da pesquisa

Para contemplar o objetivo desta revisão foram definidas as seguintes questões de pesquisa (QP):

QP 1: Quais são as formas de estruturação e publicação dos cadernos de pesquisa?

QP 2: Quais tecnologias e orientações são adotadas para publicar conjuntos de dados de cadernos de pesquisa de modo a permitir que os dados sejam encontráveis, acessíveis, interoperáveis e reutilizáveis?

c) Critérios de seleção das bases de dados

Definiu-se para este estudo selecionar as bases de dados nacionais e internacionais, com características específicas em Ciência da Informação e multidisciplinares para ampliar o escopo da recuperação de estudos primários. Definiu-se, ainda, que as bases de dados selecionadas permitam o acesso pela Comunidade Acadêmica Federada, via Portal de Periódicos Capes. As bases de dados selecionadas para busca de estudos primários são:

Quadro 03 – Seleção das Bases de Dados

Bases de dados	Características
Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)	Base de dados multidisciplinar que integra os sistemas de informações de teses e dissertações das instituições de ensino e pesquisa do Brasil.
Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI)	Indexa artigos de títulos de periódicos da área de Ciência da Informação no Brasil.
Directory of Open Access Journals (DOAJ)	Indexa títulos a nível internacional e caracteriza como área multidisciplinar.
Library and Information Science Abstracts (LISA)	Indexa títulos a nível internacional na área da Ciência da Informação.
Google Scholar	Ferramenta de busca de trabalhos acadêmicos do Google classificada como base de dados internacional e multidisciplinar.
Library, Information Science & Tecnology Abstracts (LISTA)	Indexa títulos a nível internacional na área da Ciência da Informação.
Scopus	Base de dados de resumos e citações de artigos internacionais no contexto multidisciplinar. Algumas referências apresentam o *link* de acesso ao documento.
Web of Science	Base de dados que fornece dados abrangentes de trabalhos internacionais e multidisciplinares.

Fonte**:** Elaborado pela autora (2020).

d) Critérios de estratégias de busca

Como estratégia de busca nas bases de dados, definiu-se o uso da linguagem natural, *strings* de busca e operadores booleanos (AND, OR e NO), quando necessário; consistência entre singular e plural, bem como o uso de filtros para busca e recuperação das informações. Os filtros previamente estabelecidos para busca foram o resumo, as palavras-chave e o título,

porém os filtros podem ser adaptados conforme as especificidades de cada base de dados. Os idiomas selecionados para recuperação de textos foram português, inglês e espanhol.

e) Critérios de seleção dos estudos primários (Inclusão e Exclusão)

Os critérios de inclusão definidos para a seleção dos estudos primários foram:

- Para o tema *Open Notebook Science* ou Cadernos Abertos de Pesquisa ou Cadernos de Laboratório, definiu-se considerar trabalhos publicados entre anos de 2006 a 2020, considerando o ano da introdução dos termos na literatura e o período final desta pesquisa;
- Para a *string* de busca *Open Notebook Science* AND *Semantic Web* e para todos os demais estudos, estabeleceu-se os últimos dez anos, ou seja, de 2010 a 2020;
- Fontes originais com conceitos e definições, qualquer período de publicação pode ser considerado.
- Trabalhos publicados em *slides* e *blogs*, desde que sejam compreensíveis e a autoria seja a fonte principal de informação daquele tema. Por exemplo, os *blogs* são as principais fontes de informação de Jean-Claude Bradley, o idealizador do tema *Open Notebook Science*;
- Estudos qualitativos e revisões anteriores com abordagem ampla do tema pesquisado.

Os critérios de exclusão definidos para a seleção dos estudos foram:

- Textos repetidos;
- Artigos sem disponibilidade de texto completo, via Portal de Periódicos Capes;
- Artigos fora do contexto do tema estudado, como dados governamentais, redes sociais, bibliometria, redes neurais, dentre outros;
- Trabalhos publicados na modalidade pôster, slides ou *datasets* numéricos e fórmulas, ou seja, trabalhos incompletos ou incompreensíveis;
- Textos que não estejam publicados nos idiomas inglês, espanhol e inglês;
- Textos considerados com abordagem rasa ou com apenas citações de textos de fácil acesso.

f) Critérios de seleção de estudos iniciais

Trabalhos selecionados que possuam o termo de busca no resumo e nas palavras-chave do trabalho, artigos que apresentam um resumo estruturado (objetivo, metodologia e resultado), artigos com bibliografia atualizada sobre as palavras-chave selecionadas. No caso de teses e dissertações e livros completos, o sumário foi considerado na seleção dos estudos iniciais.

g) Critérios para avaliação da qualidade

A etapa de análise da qualidade das fontes primárias visa determinar por meio de critérios bem definidos se o estudo é apropriado para responder às questões de pesquisa delimitadas no protocolo da revisão. Para este estudo, os critérios definidos foram baseados nas definições de Popay, Rogers e Willians (1998), que se referem: a) ao objeto de estudo da pesquisa, visando ao foco no contexto e ações do que é pesquisado; b) evidências de descrição adequada para interpretação do significado e contexto da pesquisa; c) adequação teórica e conceitual para descrição dos resultados e conclusões; d) qualidade dos dados com apresentação de diferentes fontes de pesquisa; e) relevância do estudo, evidenciadas por número de citações.

2.4.2 Condução da Revisão Sistemática de Literatura

Na fase de realização ou execução da revisão sistemática de literatura almeja-se fundamentalmente, encontrar documentos que sejam úteis para responder as questões da pesquisa e evitar a recuperação de documentos irrelevantes ao tema investigado. Sendo assim, a partir dos critérios previamente definidos, essa fase da pesquisa contemplou as atividades de busca e recuperação dos estudos primários nas bases de dados identificadas, seleção e análises dos documentos recuperados.

a) Buscas de Estudos Primários nas bases de dados

O processo de busca e recuperação das informações nas bases de dados foi realizado no decorrer da pesquisa desta tese e conferido nos meses de maio e junho de 2020, por meio das estratégias de busca previamente estabelecidas no protocolo de revisão. Entretanto, a estratégia de busca deve possibilitar adaptações conforme configurações de cada base de dados, o que seriam principalmente as variações de filtros e a utilização de *strings* definidas a partir de palavras-chave e palavras sinônimas. Neste estudo, os sinônimos são advindos das

diferentes traduções do termo *Open Notebook Science* para Cadernos Abertos de Pesquisa, Cadernos de Laboratório, Cadernos Eletrônicos de Laboratório e Caderno Aberto de Ciência.

Estabeleceu-se inicialmente o uso de buscas simples com *strings* constituídas por palavras-chave definidas genericamente pela expressão "open notebook science" para recuperar o maior número possível de documentos. Nas bases de dados brasileiras foi necessária realizar buscas simples pelos *strings* "cadernos abertos de pesquisa", "caderno de laboratório", "caderno eletrônico de laboratório", ou adicionar buscas combinadas a partir do uso de operador booleano OR, para a recuperação de um termo ou outro e adotar caracteres curingas, o asterisco (*), para não diferenciar entre singular e plural. Segundo Lancaster (2004), em bases de dados muito grandes torna-se difícil recuperar mais documentos úteis em relação a capacidade de evitar documentos inúteis. Sendo assim, após a primeira tentativa de busca e resposta no *Google Scholar*, optou-se por adotar a busca combinada pelos *strings* "open notebook science" AND "semantic web" e "open notebook science" AND "e-science", a qual contempla os elementos de estudo nesta tese.

Na Biblioteca Digital de Teses e Dissertações (BDTD) ocorreram as seguintes tentativas de buscas: 1) "open notebook science"; 2) "open notebook science" OR "cadern* abert* de laboratóri*" com o uso do operador booleano OR para a recuperação de um termo ou outro e caracteres curingas (*) para não diferenciar entre singular e plural; 3) "open notebook science" OR "cadern* de laboratóri*" OR "cadern* abert* de pesquisa" OR "cadern* eletrônic* de laboratório*". Considerando a baixa capacidade de recuperação de textos, estabeleceu-se o filtro 'Todos os Campos'.

Na Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (Brapci) as melhores respostas foram para as seguintes estratégias de busca: 1) "open notebook science"; e 2) "caderno eletrônico de laboratório". Nessa base, utilizou-se os filtros título, palavras-chave e resumo.

No Directory of Open Access Journals (DOAJ) foram aplicados os seguintes *strings* de busca: 1) "open notebook science"; e 2) "caderno eletrônico de laboratório". Considerando baixa quantidade de artigos indexados sobre o tema nessa fonte, manteve-se o filtro de busca com a opção 'search all' (todos os campos de pesquisa).

Na base acadêmica Google Scholar, definiu-se buscas combinadas considerando a quantidade de dados recuperados e a dificuldade de definição de filtros. As buscas ocorreram pelos *strings* "open notebook science" AND "semantic web", "open notebook science" AND "e-science" com filtros que estabelecem o recorte temporal entre 2010 e 2020 e o idioma português, inglês e espanhol.

Na Library and Information Science Abstracts (LISA), buscou-se pelo *string* “open notebook science” e selecionou-se o filtro resumo.

Na Library, Information Science & Tecnology Abstracts (LISTA), buscou-se pela *string* “Open Notebook Science” e aplicou-se o filtro ‘texto completo’ para busca e recuperação dos documentos.

Na Scopus estabeleceu-se a *string* “open notebook science” em uma busca simples e adotou-se os filtros de busca para o título, assunto e palavras-chave.

Na Web of Science adotou-se a *string* “open notebook science” e o filtro para título e tópicos.

Quadro 04 - Processo de Busca e Recuperação de Estudos Primários

Fonte	***String***	**Filtro**	**Idioma**	**Recorte temporal**	**Resposta da busca**
BDTD	"open notebook science" "cadern* eletrônic* de laboratóri*" "cadern* aberto de pesquisa" "open notebook science" OR "cadern* abert* de laboratório*" OR "cadern* abert* de pesquisa"	Todos os campos	Inglês, Espanhol Português.	2006 - 2020	1
BRAPCI	"open notebook science" "caderno eletrônico de laboratório"	Título, Palavra-chave e Resumo	Inglês, Espanhol Português	2006 - 2020	5
DOAJ	"open notebook science" "caderno eletrônico de laboratório"	Todos os Campos	Inglês, Espanhol Português	2006 - 2020	6
Google Scholar	"open notebook Science" AND "semantic web" "open notebook Science" AND "e-science"	Com todas as palavras Em qualquer lugar do artigo	Inglês, Espanhol Português	2010 - 2020	90
LISA	"open notebook Science"	Resumo	Inglês, Espanhol Português	2006 - 2020	3
LISTA	"open notebook Science"	Texto completo	Inglês, Espanhol Português	2006 - 2020	6
SCOPUS	"open notebook Science"	Título, Assunto e Palavras-chave.	Inglês, Espanhol Português	2006 - 2020	12
Web of Science	"open notebook Science"	Título e Assunto	Inglês, Espanhol Português	2006 - 2020	22

Fonte: Elaborado pela autora (2020).

Após a busca individual em cada base de dados, realizou-se a exportação dos dados nos formatos BibTEX para o software StART[1], acrônimo de *State of the Art through Systematic Review* para a conferência de documentos duplicados e seleção dos textos, a partir dos critérios de inclusão, exclusão e análise dos resumos.

b) Seleção e avaliação da qualidade dos trabalhos

Na etapa da seleção, verificou que dos 145 documentos recuperados nas bases de dados selecionadas, 39 não permitiram acesso via portal Capes/CAFe e 17 eram duplicados, os quais foram excluídos da pesquisa. Logo, passou-se a leitura dos resumos e palavras-chave dos trabalhos para aplicar os critérios de inclusão e exclusão, sendo que, dessa análise, considerou-se que 89 textos não atendiam aos critérios de inclusão estabelecidos no protocolo de revisão. Os principais critérios de exclusão dos itens foram a falta de disponibilidade de texto completo via Portal de Periódicos Capes, artigos que não apresentavam relação com o contexto dos cadernos abertos de pesquisa e suas ações, trabalhos apresentados em congressos com apenas imagens e fórmulas, tornando-os incompreensíveis para este estudo. No entanto, fez-se consultas a especialistas e autores textos, e recuperou mais 9 textos.

O processo de avaliação da qualidade tem como propósito determinar se cada trabalho identificado é apropriado à questão de pesquisa e se oferece conteúdo para sustentar o referencial teórico da pesquisa. Seguindo os critérios de qualidade definidos no protocolo da revisão sistemática, mantiveram-se os textos que apresentavam: a) caderno aberto de pesquisa como o foco dos estudos; b) adequação teórica e conceitual para descrição dos resultados e conclusões; c) fundamentação teórica e abrangência na temática, descartando-se assim os trabalhos com menos profundidade nos estudos; d) abordagem nas ações que envolvem a organização e publicação dos cadernos de pesquisa, tais como: ciclo de vida dos dados, compartilhamento de dados e as tecnologias para publicação dos dados; e) tipologias e características dos cadernos de pesquisa; e) questões regulatórias e legais.

A extração dos dados aconteceu em dois momentos, sendo que o primeiro ocorreu durante a avaliação da qualidade dos trabalhos. Nesse momento, os dados foram organizados por título, autor, data de publicação, palavras-chave e fonte de pesquisa, conforme quadro 05.

[1] StArt. Ferramenta de apoio às atividades de revisão sistemática de literatura, desenvolvida pelo Laboratório de Pesquisa em Engenharia de Software (LAPES), do Departamento de Ciência da Computação da Universidade Federal de São Carlos. Disponível em: http://lapes.dc.ufscar.br/tools/start_tool.

Quadro 05 - Fontes Primárias selecionadas para a RSL

Título	Autor	Ano	Palavras-chave	Fonte de dados
Chemistry crowdsourcing and open notebook Science	BRADLEY, Jean-Claude; OWENS, Kevin; WILLIAMS, Antony	2008	Open Notebook Science; e-Science; Web Semântica; UsefulChem	Google Scholar
A vida de laboratório: a construção de fatos científicos	LATOUR, Bruno; WOOLGAR, Steve	1997	Inscrição literária; procedimentos experimentais; contexto dos laboratórios.	Indicação de especialista
Open Notebook Science using blogs and wikis.	BRADLEY, Jean-Claude	2007	Blogs; wikis; publicação de cadernos de pesquisa; UsefulChem.	Indicação de especialista
Pseudo open notebook science?	BACON, Dave	2008	Open Notebook Science; Logos de abertura	Indicação de especialista
Open notebook science claims and logos	BRADLEY, Jean-Claude	2009	UsefulChem; indicação de logos de uso	Indicação de especialista
Open notebook science, reproducibility and exclusion.	BRADLEY, Jean-Claude	2009	UsefulChem	Google Scholar
The Impact of Open Notebook Science	POYNDER, Richard	2010	Entrevista Jean-Claude Bradley; Open Notebook Science; UsefulChem; Contexto histórico	LISA e LISTA
e-Research in the life sciences: from invisible to virtual colleges	POWER, Lucy	2011	Open Laboratory Notebooks (OLNs); Open Notebook Science	Google Scholar
Resource description framework technologies in chemistry	WILLIGHAGEN, Egon; BRANDLE, Martin	2011	Web Semântica, RDF, Ontologias	Google Scholar
Mining chemical information from open patentes	JESSOP, David M; ADAMS, Sam E; MURRAY-RUST, Peter	2011	Web Semântica, tecnologias, Patentes	Google Scholar
Colaboration using open notebook science in academia. In: Collaborative computational technologies for biomedical research.	BRADLEY, Jean-Claude; LANG, Andrew; KOCH, Steve; NEYLON, Cameron	2011	UsefulChem; Wikis; publicação de experimentos; propósitos dos cadernos	Google Scholar
Open source drug discovery for malaria. The Synaptic Leap open source biomedical research	TODD, Matthew	2011	Open Source Malaria; Definições de abertura	Indicação de especialista
Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences	BIRD, Colin L; FREY, Jeremy G.	2013	Open Notebook Science e-Science	Google Scholar, Scopus e Web of Science
Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences	BIRD, Colin L.; WILLOUGHBY, C.; FREY, Jeremy G	2013	Características e usabilidade de cadernos de laboratório; UsefulChem	Google Scholar
Jean-Claude Bradley: hero of open notebook science; it must	MURRAY-RUST, Peter	2014	Jean-Claude Bradley; homenagem a	Indicação de

become the central way of doing science.			Bradley	especialista
10	ZHU; PROCTER	2015	Mídias sociais, comunicação acadêmica, blog, Twitter, Facebook, Web 2.0, PhDestudantes, pesquisadores em início de carreira	Google Scholar
Facilitating chemical discovery: an e-science approach	MILSTED, Andrew J.	2015	Repositórios digitais, lista de open notebooks,	Google Scholar
Ten simple rules for a computational biologist´s laboratory notebook.	SCHNELL, S.	2015	Caderno de laboratório;	Indicação de especialista
Novos cadernos de laboratório e novas culturas epistêmicas: entre a política do experimento e o experimento da política	CLINIO, Anne	2016	Cultura epistêmica, Caderno de laboratório, Open Notebook Science, Caderno aberto de laboratório, Laboratório cidadão, Ciência aberta, ciência comum	BDTD
Open notebook science as an emerging epistemic culture within the Open Science movement	CLINIO, Anne; ALBAGLI, Sarita	2017	Open notebook science, Jean-Claude Bradley, Epistemic culture, Matter of proof.	DOAJ e Google Scholar
Cadernos abertos de laboratório e publicações líquidas: novas tecnologias literárias para uma Ciência Aberta	CLINIO, Anne; ALBAGLI, Sarita	2017	Open Notebook Science, Caderno aberto de laboratório, Ciência aberta, Jean-Claude Bradley, Cultura epistêmica, Publicação liquida.	Brapci e DOAJ
Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa	ROCHA; SALES; SAYÃO	2017	Cadernos eletrônicos de laboratório, dados de pesquisa, ciência aberta, preservação de dados, acesso aberto	Brapci
Open notebook science can maximize impact for rare disease projects	HARDING, Rachel J.	2019	LabScribbles, OpenLabnotebooks	Web of Science, Scopus e DOAJ
Documentos de arquivo produzidos pela atividade científica: uma análise dos cadernos de laboratório do Instituto Oswaldo Cruz	SANTOS; BORGES; LORENÇO	2019	documento de arquivo; caderno de laboratório; arquivo de laboratório; diagnóstico de arquivo	DOAJ
Open laboratory notebooks: good for science, good for society, good for scientists	SCHAPIRA, Matthieu; HARDING, Rachel	2019	Caderno OpenLabnotebooks	Google Scholar
Why should you keep na open notebook?	OPEN NOTEBOOK SCIENCE NETWORK	2019	Open Notebook Science; sociedade científica	Indicação de especialista
Protocol data management in biology laboratories: proposal for the development of na information management system.	FARIA-CAMPOS, Alessandra C. *et al*.	2020	Protocolo de pesquisa de biologia; organização do conhecimento	Indicação de especialista

Fonte: Elaborado pela autora (2020).

No segundo momento, realizou-se a extração dos dados levando em consideração a classificação dos estudos primários por assunto. Nessa classificação, definiu-se os aspectos temáticos dos trabalhos seguidos dos autores e ano de publicação.

Quadro 06 - Fontes Primárias classificadas por abordagens

Abordagens	Autores
Caderno LabScrilbes	HARDING (2016), HARDING (2019)
Caderno Open Source Malaria	ABDO (2014), ALBAGLI; CLINIO; RAYCHTOCK (2014)
Caderno OpenLabnotebooks	SCHAPIRA; HARDING (2019), HARDING (2019)
Caderno RedLab	BIRD; FREY (2013)
Caderno UsefulChem	BRADLEY (2007), BRADLEY (2009), BRADLEY; OWENS; WILLIAMS (2008), CAMERON (2008), CAMERON, LYON (2009), BRADLEY (2010), POYNDER (2010), BRADLEY; LANG, KOCH; NEYLON (2011), POWER (2011), BIRD; FREY (2013), BIRD; WILLOUGHBY; FREY (2013), ABDO (2014), CLINIO (2016); CLINIO; ALBAGLI (2015)
Características dos tipos de cadernos de pesquisa	ROCHA; SALES; SAYÃO (2017)
Ciclo de Vida dos dados	KWALCZYK; SHANKAR (2011), BIRD; FREY (2013)
Compartilhamento de dados	KWALCZYK; SHANKAR (2011)
Dados de pesquisa científica: tipologias, princípios e diretrizes, gestão e curadoria de dados e compartilhamento de dados.	ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2007), BORGMAN (2010), KWALCZYK; SHANKAR (2011), TENOPIR (2011; 2015), FORCE11 (2014), BIRD; FREY (2013), APPEL (2014), SAYÃO; SALES (2015), SANTA'ANA (2016), FERREIRA (2018), SILVA (2019).
Diretrizes semânticas para publicação de cadernos de pesquisa	Não identificado.
Fundamentação teórica e abrangência sobre cadernos abertos de pesquisa	BRADLEY; OWENS; WILLIAMS (2008), POYNDER (2010), POWER (2011), ALBAGLI; CLINIO; RAYCHOTCK (2014), CLINIO (2016), CLINIO; ALBAGLI (2017a), CLINIO; ALBAGLI (2017b)
Infraestrutura da *e-Science*, tecnologias da Web Semântica e *Linked Data* relacionada a publicação de dados	GRAY (2009), HEY; TANSLEY; TOLLE (2009), SANTAREM SEGUNDO (2014; 2015; 2018), CONSÓRCIO W3C, APPEL (2014), LAUFER (2015), SCHAPIRA (2019), ISOTANI; BITEENCOURT (2019), RAUTENBERG; SOUZA; DALL'AGNOL; MICHELON (2018)
Natureza dos dados	KWALCZYK; SHANKAR (2011)
Princípios FAIR – Caderno aberto de pesquisa	Não identificado
Produção documental em Laboratório de pesquisa	SANTOS; BORGES; LOURENÇO (2019)
Publicação de dados de cadernos abertos de pesquisa	ZHU; PROCTER (2015), ERKIMBAEV *et al*. (2013), FARIA-CAMPOS *et al*. (2020)
Questões regulatórias e legais de cadernos de pesquisa	BIRD; WILLOUGHBY; FREY (2013)
Tecnologias da Web Semântica	ERKIMBAEV *et al*. (2013), WILLIGHAGEN; BRANDLE (2011); JESSOP; ADAMS; MURRAY-RUST (2011)
Vantagens e desvantagens dos cadernos de pesquisa	ROCHA; SALES; SAYÃO (2017)
Web Semântica e Caderno aberto de pesquisa	ERKIMBAEV *et al*. (2013); WILLIGHAGEN; BRANDLE (2011), JESSOP; ADAMS; MURRAY-RUST (2011)

Fonte: Elaborado pela autora (2020).

2.4.3 Resultado da Revisão Sistemática de Literatura

Essa etapa do estudo apresenta os resultados da RSL quanto às questões elaboradas no protocolo da pesquisa. O retorno obtido sobre as questões **QP1** (Quais são as formas de estruturação e publicação dos cadernos de pesquisa?) e **QP 2** (Quais tecnologias e orientações são adotadas para publicar conjuntos de dados de cadernos de pesquisa de modo a permitir que os dados sejam encontráveis, acessíveis, interoperáveis e reutilizáveis?) consiste em:

As buscas na literatura sobre as formas de publicação de cadernos de pesquisa retornaram os registros desde caderno de papel, *blog*, *wikis*, cadernos eletrônicos (fechados) até os cadernos abertos de pesquisa (a partir do conceito de Jean-Claude Bradley, objeto desta tese) publicados em repositórios de dados.

Bird, Willoughby e Frey (2013) apresentam o contexto evolutivo dos cadernos de laboratório de 1980 até 2010, iniciando com a abordagem do caderno de papel como uma maneira de capturar e preservar os registros das atividades humanas, científicas e outras. Os cadernos de papel têm as vantagens de serem portáteis, não requerem fonte de alimentação e podem ser armazenados com segurança. As desvantagens são os riscos de perda ou destruição e as complicações com a recuperação de material, principalmente dados, para reutilização. Nesse contexto, a Collaborative Electronic Notebook Systems Association (CENSA) (1998) – um consórcio dedicado a promover sistemas de manutenção de registros eletrônicos e tecnologias colaborativas, inativo desde 2016, forneceu uma lista de razões pelas quais os cadernos de papel são considerados obsoletos. Como esse recurso não é objeto de estudo desta tese, o tema não será aprofundado. Os autores Bird, Willoughby e Frey (2013) argumentam a necessidade de soluções que cubram o meio termo entre arquivos simples não controlados e os sistemas de gerenciamento relativamente inflexíveis, sugerindo que a Web Semântica seja a alternativa necessária para o ambiente do laboratório.

Outra forma recorrente de publicação de cadernos de pesquisa são as redes sociais. Para Bradley, Owens, Williams (2008), os seus projetos são hospedados gratuitamente usando funções gerais de *blogs* e *wikis* para facilitar a replicação em quaisquer domínios científicos, mas garantem que esses serviços não são quimicamente inteligentes e se limitam apenas ao compartilhamento de dados baseados em texto e gráfico. No entanto, já vislumbravam o uso tecnologias semânticas em um movimento da Ciência Aberta com o desejo de publicar o processo científico de forma mais transparente e com mais eficácia. Esse processo não só permitiria o compartilhamento contínuo de dados, mas se propunha a oportunidade de desenvolver comunidades colaborativas entre cientistas. Os autores Zhu e Procter (2015)

complementam que as comunicações de cadernos abertos de pesquisa são frequentemente publicadas em *blogs*, *Twitter*, *facebook*, dentre outras redes sociais.

A tese denominada *e-Research in the Life Sciences*, de autoria de Power e College (2011), estuda a utilização de ferramentas on-line na área da Ciência da Vida e constata que a utilização de ferramentas como redes sociais, cadernos de laboratório e *blogs* de ciência tem impacto significativo no movimento científico. Esta tese faz menção ao uso dos metadados (título, autor e fonte) no contexto de *blogs*.

Para Bradley, Lang, Koch e Neylon (2011), o uso de *wiki* para um caderno de laboratório também se tornou conveniente para a comunicação entre pesquisadores e alunos. Ainda segundo Bradley, Lang, Koch e Neylon (2011), o *Wikispaces* foi a plataforma escolhida para o projeto, pois oferece um serviço gratuito hospedado para *wikis* públicos e oferece editor visual intuitivo, *wikitext* simplificado e recursos convenientes de *backup* e alerta. Esse contexto apresentado é histórico, visto que em outras citações foi possível verificar a evolução e mudanças das plataformas adotadas pelos autores, conforme apresentado na revisão de literatura sobre Cadernos Abertos de Pesquisa.

Ao tratar dos cadernos de pesquisa Bird e Frey (2013) consideram os conceitos de preservação, curadoria, proveniência, descoberta e acesso no contexto do ciclo de vida da pesquisa, e enfocando no papel dos metadados, particularmente as ontologias das quais a Web Semântica da química emergente dependerá. No entanto, Bird e Frey (2013) não apresentam uma estrutura de metadados e vocabulários para a descrição dos dados gerados nos laboratórios de química, área de estudo dos autores.

Willighagen e Brandle (2011) não tratam especificamente de cadernos abertos de pesquisa, mas descrevem tecnologias da Web Semântica indicadas para a área na química e exemplifica que Murray-Rust adotou o modelo RDF em seu projeto de caderno eletrônico de laboratório.

Como pode ser observada, a tendência para a publicação de cadernos abertos de pesquisa refere-se à abordagem que envolve as tecnologias da Web Semântica. Os trabalhos mencionaram a importância e as vantagens da aplicação de tecnologias semânticas, uso de metadados de proveniência e preservação, e vocabulários padronizados, mas não foi possível identificar como esses elementos se estruturam para a implementação em plataformas digitais. Nesse sentido, volta-se ao objetivo da revisão sistemática que é identificar lacunas no conhecimento, aponta-se a necessidade de apresentar como são estruturados os dados e metadados para descrição, preservação e proveniência dos dados que compõem o ecossistema dos cadernos abertos de pesquisa, pois não basta informar que os repositórios adotados são

semânticos. Acredita-se haver a necessidade de explicitar e documentar toda a estruturação por trás da plataforma digital para colaborar com outros projetos. Sendo assim, é nessa direção que esta tese seguirá, ou seja, apresentar um conjunto de diretrizes semânticas para estruturação e publicação de dados de pesquisa científica anotados em cadernos abertos de pesquisa com vista a colaborar com a qualidade da recuperação da informação.

2.5 ESTRUTURA DA TESE

Esta pesquisa foi organizada em quatro partes e subdividida em capítulos, conforme apresentados a seguir:

1ª Parte – constituída por elementos introdutórios e metodológicos da pesquisa.

Capítulo 1 – Introdução - este capítulo contextualizou o tema da pesquisa à luz da Ciência da Informação. Na justificativa foram analisadas pesquisas com temas correlatos para identificar estudos na temática, a relevância e a motivação desta tese. Ainda, neste capítulo, foram apresentadas a questão problema, a hipótese construída e os objetivos da pesquisa.

Capítulo 2 – Metodologia e Estrutura da Tese – o capítulo apresenta os aspectos metodológicos, incluindo universo e amostra da pesquisa, classificação da pesquisa, coleta de dados e revisão sistemática de literatura, além da estrutura desta tese.

2ª Parte – constituída pelos capítulos que compõem a revisão de literatura, eixos temáticos propostos para viabilizar a pesquisa.

Capítulo 3 – Dimensões da *e-Science* – apresenta os aspectos conceituais e históricos da *e-Science*, bem como as suas características, funções e fatores que levam à formulação do quarto paradigma científico. A contextualização desse capítulo é fundamental para compreender o desdobramento desta pesquisa no atual cenário da comunicação científica.

Capítulo 4 – Dados de Pesquisa Científica – descreve os dados de pesquisa científica em suas abordagens conceituais e tipológicas. Apresenta modelos de ciclo de vida dos dados, com ênfase nos modelos *Curation Lifecycle Model* (*DCCCuration*) e o *DataONE Data Lifecycle* como instrumentos para gerenciamento de dados de pesquisa; e princípios para o gerenciamento e publicação de dados de pesquisa na Web com destaque para os *FAIR Principles* [Princípios FAIR], publicado pela FORCE11(2014) e *Data on the Web Best Practices* [Melhores Práticas para Publicação de Dados na Web], publicadas pelo Consórcio W3C (2017).

Capítulo 5 – Cadernos Abertos de Pesquisa – apresenta os cadernos abertos de pesquisa em seus aspectos conceituais e históricos. Contempla iniciativas que mostram ações que vem sendo desenvolvidas no contexto da abertura de dados de pesquisa. Faz-se uma

análise da estrutura dos dados anotados nos cadernos de pesquisa a partir dos modelos de cadernos apresentados no capítulo.

Capítulo 6 - Web Semântica e *Linked Data* – aborda a Web Semântica a partir de suas tecnologias e os conceitos do *Linked Data,* com ênfase nos vocabulários e orientações para publicação de dados de pesquisa científica. Para definir as etapas das diretrizes serão analisadas tecnologias recomendadas nas principais orientações para dados na Web, apresentadas nos capítulos anteriores, como as Melhores Práticas para dados na Web e os Princípios FAIR.

3ª Parte – constituída pela discussão e formulação do modelo proposto.

Capítulo 7 – Conjunto de Diretrizes Semânticas para Publicação de Cadernos Abertos de Pesquisa - apresenta as etapas e elementos identificados para compor as diretrizes semânticas para estruturação e publicação de cadernos abertos de pesquisa. Para a modelagem dos dados, adotou-se o modelo conceitual IFLA *Library Reference Model* (IFLA LRM) (2017). Destacam-se a identificação do ecossistema dos dados de pesquisa científica em torno dos cadernos de laboratório, mapeamento e descrição dos metadados, mapeamento dos vocabulários, classes e propriedades correspondes aos metadados. Ainda nessa seção realizou-se uma análise dos elementos semânticos estabelecidos quanto aos Princípios FAIR e melhores práticas recomendadas pelo Consórcio W3C.

Capítulo 8 – Considerações Finais, Desafios Enfrentados e Perspectivas Futuras

4ª Parte – constituída pela apresentação dos elementos pós-textuais à pesquisa.

Capítulo 9 – Referências.

3 E-SCIENCE

A *e-Science* é compreendida como um conjunto de ferramentas e tecnologias de alto desempenho para apoiar a pesquisa cientifica intensiva em dados, multidisciplinar e colaborativa em rede (HEY;TANSLEY; TOLLE, 2009). As principais características que assinalam a *e-Science* são o enorme volume de dados de pesquisa, a exploração da tecnologia de larga escala e a colaboração entre pesquisadores de diversas áreas do conhecimento.

O termo *e-Science* é traduzido para o português como e-Ciência, o qual adquiriu um significado que representa a potência da ciência melhorada com o uso intensivo das tecnologias da informação e comunicação (TICs) e sua aplicação em torno de um esforço colaborativo (FERREIRA, 2018, p. 2). Para Vaz (2011, p. 20) o ‘e’ de *e-Science* iniciou com o significado de eletrônico (*electronic*) – ciência eletrônica - e agora representa melhor (*enhanced*) ou habilitada (*enabled*). Nesse caso, a ciência melhorada vem trazendo grandes avanços para o desenvolvimento científico e modernizando a ciência frente ao contexto atual.

Na literatura, a *e-Science* se apresenta com diferentes terminologias, como o quarto paradigma (*fourth paradigma*), ciência orientada por dados (*data-driven Science*), computação fortemente orientada a dados (*data-intensive computing*), mineração de dados (*data mining*), pesquisa eletrônica (*e-Research*) e dos dados ao conhecimento (*from data to knowledge*) (CESAR JUNIOR, 2011). Desses termos, destaca-se o quarto paradigma (*fourth paradigma*), que é tratado neste estudo como parte da evolução da ciência que marcou o surgimento e as características que assinalam o fenômeno da *e-Science.*

Outro termo adotado com o mesmo propósito da *e-Science* é o ciberinfraestrutura (*cyberinfrastructure*), porém seguimos a linha de entendimento que a ciberinfraestrutura denota uma das infraestruturas que compõem a *e-Science*, sendo uma condição para o desenvolvimento e realização da *e-Science*. Além desse termo, ressalta-se o fenômeno de geração de grandes volumes de dados que vem sendo identificado como *Big Data.* Esse movimento produz dados que excedem as capacidades convencionais de processamento dos sistemas de bases de dados, os quais podem variar em volume, velocidade de crescimento, variabilidade, e vem se ocupando de outros campos e com finalidades que extrapolam a pesquisa científica (APPEL, 2014). Para o autor, no campo da ciência, a terminologia *e-Science* é mais comumente utilizada e possui fatores preponderantes que os diferencia como a pesquisa colaborativa e o uso de recursos compartilhados para a exploração de dados.

Sendo assim, neste estudo, adota-se o termo *e-Science* para ressaltar que a partícula ‘e’ representa a ciência melhorada e atuante no uso de dados de pesquisa científica por meio de métodos computacionais sofisticados e computação avançada.

O fenômeno *e-Science*, cenário do movimento da Ciência Aberta, vem impulsionando novas práticas de produção do conhecimento, sobretudo, no que se refere à publicação semântica de dados de pesquisa científica, proporcionando o uso e reuso de dados em novos contextos, sem que haja má interpretação destes e proporcionando novas descobertas.

As dimensões da *e-Science* definidas, neste estudo, compostas por Tecnologias, Dados de Pesquisa Científica e Colaborações Científicas em Rede podem potencializar a publicação de dados registrados em cadernos de pesquisa durante todas as atividades de uma pesquisa.

Nesse sentido, este capítulo buscou identificar os elementos conceituais e práticos presentes nas dimensões *e-Science* que precisam ser observados para a publicação de dados de pesquisa científica. Sendo assim, as abordagens teóricas indispensáveis à compreensão do tema foram: 3.1 evolução dos paradigmas da ciência e 3.2 dimensões da *e-Science*.

Na seção 3.1 é apresentada a evolução dos paradigmas da ciência, buscando conhecer as características que marcaram a origem e o desenvolvimento da *e-Science*, destacando as principais evidências conceituais definidas pelos autores da área, para então identificar as dimensões e seus principais elementos que devem compor o modelo de publicação de dados de pesquisa. Os elementos foram agrupados nas dimensões Tecnológica, Dados de Pesquisa Científica e Colaboração Científica em Rede. Tais dimensões devem ser aplicadas de forma inter-relacionadas e interdependentes para possibilitar a sistematização do ciclo de vida dos dados de pesquisa.

Na seção 3.2 são discutidas as dimensões tecnológica, dados de pesquisa científica e colaboração científica em rede, a partir do modelo em camadas de infraestrutura eletrônica da *e-Science*, definido por Andronico e colaboradores (2011), sendo que a dimensão tecnológica contempla as infraestruturas computacionais de uso distribuído necessárias para a estruturação dos dados de pesquisa e colaboração entre pesquisadores; a dimensão dados de pesquisa científica refere-se ao exorbitante volume de dados gerados a partir da convergência das tecnologias de comunicação e computação; e dimensão colaboração, trata-se da interação multidisciplinar entre cientistas na produção de dados, independente de fronteiras geográficas por meio de comunidades virtuais.

3.1 EVOLUÇÃO DO QUARTO PARADIGMA DA CIÊNCIA

O termo *e-Science* foi introduzido por John Taylor, no final dos anos 90[2], no Reino Unido, para se referir ao uso intensivo de tecnologias como apoio ao desenvolvimento de pesquisas científicas colaborativas. Na ocasião, Taylor ocupava o cargo de Diretor Geral dos Conselhos de Pesquisa do Escritório de Ciência e Tecnologia do Reino Unido (OST) e, anteriormente, o cargo de chefe dos Laboratórios Europeus de Pesquisa da Hewlett-Packard (HP). A partir de suas experiências no ramo da pesquisa, Taylor observou que muitas áreas da ciência estavam se tornando cada vez mais dependentes de novas formas de trabalho colaborativo e multidisciplinar, e então, denominou esse movimento de *e-Science.* Para Taylor, a "*e-Science* trata da colaboração global entre áreas-chave da ciência e o uso da nova geração de infraestrutura para possibilitar essas novas formas de trabalho" (HEY; TREFETHEN, 2002, p. 1018; 2003, p. 1809, tradução nossa).

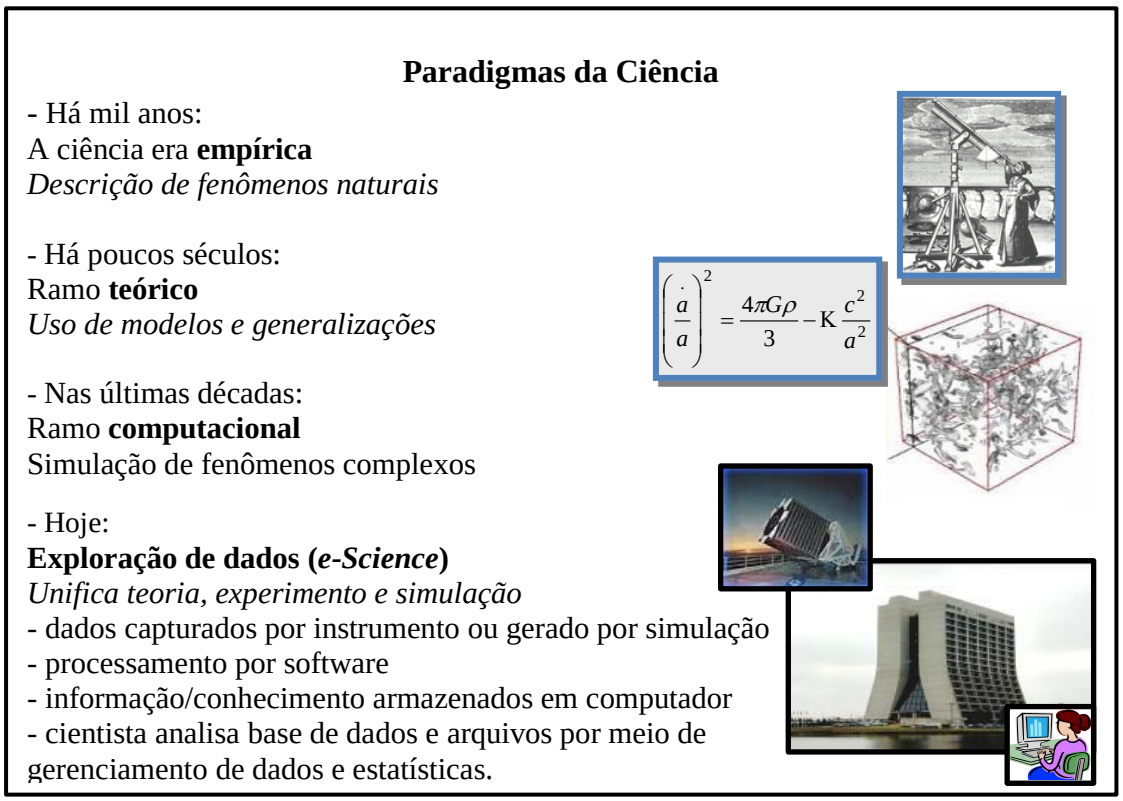
A *e-Science* surge no contexto da evolução da Ciência, caracterizando-se como um novo paradigma. Na obra 'A estrutura das revoluções científicas', de Kuhn (1998, p. 219), o termo paradigma é tratado de forma intrinsecamente circular, sendo "aquilo que os membros de uma comunidade partilham e, inversamente, uma comunidade científica consiste em homens que partilham um paradigma". Ainda para Kuhn (1998), o paradigma científico é entendido como um marco da história da epistemologia e da própria ciência. Assim, pode-se compreender que um novo paradigma da ciência representa um marco histórico que emerge das circunstâncias do momento e que são partilhados entre a comunidade científica. No momento, uma das principais circunstâncias que emergem e impulsionam o movimento da *e-Science* é a disponibilidade eminente de grandes quantidades de dados decorrentes das novas gerações de experimentos e pesquisas científicas (HEY; TREFETHEN, 2003).

Para Gray e Szalay (2007), a *e-Science* é considerada o quarto paradigma da ciência. O primeiro paradigma, mil anos atrás, foi marcado pela descrição dos fenômenos naturais, em que a Ciência era eminentemente empírica. No segundo paradigma, por volta do século XVII, a ciência se definiu a partir de modelos teóricos como as Leis de Kepler, as Leis do movimento de Newton, as equações de Maxwell e assim por diante, assumindo um ramo teórico. Na metade do século XX, a ciência passa a valer-se de uma vertente computacional, com forte apelo à pesquisa por meio da simulação de fenômenos complexos, trazendo

[2] A data da introdução do termo *e-Science* não é consenso entre os autores. De acordo com a The Digital Curation Centre (DCC) - do português Centro de Curadoria Geral – a *e-Science* foi criada em 1999. Disponível em: http://www.dcc.ac.uk/digital-curation/glossary#E. Para Appel (2014) o termo foi introduzido em 2001. E tantos outros autores apenas se referem ao período em que Taylor ocupava o cargo de diretor na OST.

soluções de problemas complicados de resolver analiticamente com os modelos teóricos. Segundo Hey e Hey (2006), a simulação é usada para explorar domínios inacessíveis às metodologias tradicionais de pesquisa, a qual passou a gerar muitos dados juntamente com um grande aumento de dados das ciências experimentais. Na interpretação de Ferreira (2018), a computação passou a ditar as regras seguintes e a conduzir de modo consistente os rumos da Ciência, designando assim o terceiro paradigma da Ciência. Na atualidade, a evolução das tecnologias computacionais vem proporcionando a produção de grandes volumes de dados, em que se busca a interlocução de teorias, experimentos e simulações, caracterizando o momento da ciência de uso intensivo em dados ou nova era da ciência orientada a dados, denominada de *e-Science* ou quarto paradigma da ciência (GRAY, 2009). A figura 02 apresenta, na visão de Gray e Szalay (2007), a evolução dos paradigmas da ciência.

Figura 02 - Paradigmas da Ciência



Fonte: Gray e Szalay (2007, *slides*, tradução nossa).

A trajetória da ciência se apresenta sem que ocorram rupturas de paradigmas, e sim "uma confluência e aprimoramento de teorias, métodos, modelos, práticas e funcionalidades diferentes e revalidadas" (OLIVEIRA; SILVA, 2016, p. 10). Nesse cenário, observa-se que a convergência das tecnologias de comunicação e computação, especialmente o

desenvolvimento de comunidades virtuais de pesquisa, vem impulsionando as mudanças na maneira de fazer e comunicar a ciência. O movimento do quarto paradigma tende a facilitar novas dimensões de pesquisa e experimentação por meio da colaboração interdisciplinar global, envolvendo pessoas e recursos, sem que haja fronteiras geográficas.

Na busca pela compreensão conceitual do quarto paradigma da ciência foram identificadas abordagens que se assemelham e evidenciam aspectos que tendem a caracterizar o momento atual. Nesse sentido, destacam-se os dizeres de Castells (2006) ao comparar os fatos que marcaram as revoluções industriais com as transformações atuais.

> O cerne das transformações que estamos vivendo no momento atual refere-se às tecnologias da informação, processamento e comunicação. A tecnologia da informação é para esta revolução o que as novas fontes de energia foram para as revoluções industriais sucessivas, do motor a vapor à eletricidade, aos combustíveis fósseis e até mesmo à energia nuclear. (CASTELLS, 2006, p. 68).

As revoluções industriais causaram aceleração sem antecedentes históricos, modificando a economia e a sociedade, por exemplo, no setor de tecelagem, em que os primeiros teares movidos a vapor produziam vinte vezes a mais que um tear manual, consequentemente, as primeiras máquinas para fiar movidas à energia podiam produzir duzentas vezes a capacidade dos teares a vapor (CASTELLS, 2006). Hoje, algumas áreas da ciência estão enfrentando aumentos de centenas a milhares de vezes nos volumes de dados de satélites, telescópios, instrumentos de alta velocidade, redes de sensores, aceleradores e supercomputadores, em comparação com os volumes gerados apenas há uma década (HEY; TANSLEY; TOLLE, 2009).

Do ponto de vista tecnológico, a computação é considerada por Castells (2006) como elemento meio de transformação capaz de impulsionar a construção colaborativa do conhecimento, pois afirma que as aplicações tecnológicas vêm progredindo a passos gigantescos, tornando-se cada vez mais dispendiosas e melhores, com isso possibilitando sua aquisição e utilização em larga escala. De forma geral, Castells (2006) discute que os processos dominantes na era da informação estão cada vez mais organizados em torno de redes. Para Castells (2006), redes são estruturas abertas capazes de expandir de forma ilimitada, integrando novos nós desde que consigam comunicar-se dentro da rede, ou seja, desde que compartilhem os mesmos objetivos. Marteleto (2001, p. 72) expõe que rede é um "sistema de nodos e elos; estrutura sem fronteiras; uma comunidade não geográfica; um sistema de apoio ou um sistema físico que se pareça com uma árvore ou uma rede." Segundo Tomaél (2008), as redes de conhecimento agregam pessoas por objetivos comuns,

compartilham informações e possibilitam um ambiente mais produtivo na construção do conhecimento.

Ao enfatizar o uso de tecnologias e o trabalho colaborativo na ciência, Hey e Hey (2006) argumentam que muitas áreas da ciência estariam prestes a ser transformadas pela disponibilidade de vastas quantidades de novos dados científicos que poderiam potencialmente fornecer *insights* com um nível de detalhe nunca antes visto. Para tanto, a *e-Science* seria a revolução necessária para suportar a ciência baseada em dados e em rede. Portanto, definiram a *e-Science* como sendo "um conjunto de ferramentas e tecnologias necessárias para apoiar a ciência colaborativa em rede" (HEY; HEY, 2006, p. 516).

Gray e Szalay (2007) e Gray (2009) apontam a necessidade de melhorias na produção de ferramentas para suportar e sistematizar o ciclo da pesquisa, desde a captura e curadoria dos dados à análise e visualização de dados. Os autores relatam que as ferramentas atuais são frágeis e não possibilitam a curadoria e análise de dados. Por isso, muitos dados são coletados, mas não selecionados ou publicados de forma sistemática. Sendo assim, "a *e-Science* é onde a TI encontra os cientistas" (GRAY, 2009, p. 18), ou seja, os pesquisadores demandam por infraestruturas tecnológicas adequadas para desenvolverem ações de gestão de dados científicos. Além disso, o momento atual da ciência se volta para a valorização do compartilhamento dos dados de pesquisa (GRAY, 2009).

Ainda reforçando o aspecto tecnológico, Ribes e Lee (2010, p. 232) definem a *e-Science* como "um nome dado para as tecnologias de informação de apoio a atividades de investigação científica, como a colaboração no compartilhamento de dados e divulgação de resultados". Para Medeiros e Caregnato (2012), o componente humano é essencial para o desenvolvimento da *e-Science*. Dessa forma, destacam três aspectos definidos por Ribes e Lee (2010) que tendem a caracterizar as transformações proporcionadas pela *e-Science*: (a) a comunidade de colaboração ampla e interdisciplinar; (b) a coleta, representação e análise de dados dirigidos computacionalmente; e (c) a integração *end-to-end*, ou seja, integração entre pesquisadores. Para Appel (2014, p. 11), a *e-Science* é "um momento de exploração de grandes quantidades de dados derivados de uma ampla variedade de tecnologias e da colaboração entre pessoas, realizada em larga escala". Ferreira (2018, p. 13) complementa que *e-Science* é o resultado da "necessidade de uma tecnologia adicional que suportasse o desenvolvimento das pesquisas e aliviasse o isolamento do pesquisador; um composto de hardware, software e um preponderante cunho colaborativo".

Pode-se notar que as características essenciais para a implementação de projetos *e-Science* estão compostas por três fatores: (a) a construção de uma infraestrutura

computacional para o uso distribuído ou para processamento de larga escala; (b) a produção e o uso intensivo de dados; e (c) a colaboração entre cientistas, grupos de cientistas ou instituições, pelo compartilhamento de esforços, dados e/ou recursos computacionais (APPEL, 2014, p. 12). Segundo Oliveira e Silva (2016), além desses fatores a *e-Science* apresenta características aprimoradas da ciência tradicional, tais como: compartilhamento, colaboração, preservação e a atribuição autoral.

As evidências conceituais apresentadas, nesta seção, possibilitaram a identificação dos principais elementos que compõem a *e-Science*. Para a publicação de dados de pesquisa, esses elementos foram agrupados didaticamente em dimensões facetadas que são aplicadas de forma inter-relacionadas e interdependentes para possibilitar a sistematização do ciclo de vida dos dados de pesquisa, sobretudo no que se refere à publicação semântica de dados de pesquisa científica.

3.2 DIMENSÕES DA *e-SCIENCE*

Segundo a Association of Research Library por meio da Joint Task Force on Library Support for e-Science (2007, p. 6), "a *e-Science* requer novas estratégias de suporte à pesquisa e significante desenvolvimento de infraestrutura". Nesse sentido, Andronico *et al*. (2011) referem-se a *e-Science* como um método científico que prevê a adoção de plataformas digitais conhecidas como e-Infraestrutura (*e-Infrastrucuture*) que traz em seu contexto recursos que envolvem todo o processo da pesquisa, desde a concepção da ideia até a produção do resultado científico. Dessa forma, as dimensões definidas neste estudo referem-se ao conjunto de elementos que compõe a infraestrutura para o desenvolvimento de projetos *e-Science*.

No que se refere aos elementos que compõem a *e-Science*, Andronico *et al*. (2011) apresentam as e-Infraestruturas conceitualmente em três camadas, a saber:

1. Parte inferior, composta pelos instrumentos científicos e experimentais que fornecem grande quantidade de dados;
2. Em seguida, a camada de rede, centros de processamento de dados em rede e *software middleware* como a "cola" dos recursos; e
3. O terceiro e mais alto nível inclui pesquisadores que realizam suas atividades independentes da localização geográfica, interagem com os colegas, compartilham e acessam os dados (ANDRONICO *et al*. 2011, p. 156).

O modelo apresentado por Andronico *et al*. (2011) foi representado por Ferreira (2018) em formato de pirâmide, o qual permitiu melhor visualização das camadas.

Figura 03 - Modelo de Infraestrutura da e-Science

Fonte: Elaborada por Ferreira (2018, p. 22), a partir de Andronico *et al*. (2011, p. 156).

Os instrumentos científicos adotados em laboratórios de pesquisa para realização de experimentos e simulações estão se tornando cada vez mais complexos e produzindo enormes quantidades de dados, em formato digital. Esses dados geralmente são relativos às análises interdisciplinares e precisam ser analisados por comunidades cada vez maiores de pesquisadores, cujos membros estão distribuídos por todo o mundo e pertencem a diferentes domínios geográficos, administrativos, científicos e culturais (ANDRONICO *et al*., 2011). As novas práticas científicas baseadas em grandes volumes de dados recebem o reconhecimento da National Science Foundation (NSF) (2015) ao mencionar que os dados digitais não são meramente resultados de pesquisas, armazenados e acabados, mas fornecem informações para novas hipóteses, possibilitando novos *insights* científicos e impulsionando a inovação.

Na camada do meio encontra-se a infraestrutura computacional de uso distribuído. Segundo Andronico *et al*. (2011), a tecnologia *Grid* e a rede subjacente constituem o que é comumente chamado de infraestrutura eletrônica. A computação em *Grid* "potencializa o uso de computadores em rede, capaz de permitir a colaboração à distância de equipes de pesquisa, envolvendo o uso intensivo e o compartilhamento de dados e recursos computacionais" (ALBAGLI; APPEL; MACIEL, 2014, p. 2). Andronico *et al*. (2011) explicam que a tecnologia *Grid* é dispositivo de computação e armazenamento, ligados entre si por redes de banca larga, no qual o software *middleware* é instalado, permitindo que os recursos se comportem como um único computador distribuído na estrutura da internet e pode ser acessado onipresentemente por meio de serviços virtuais e interfaces de usuários de alto nível.

No topo da infraestrutura, encontram-se “as práticas colaborativas contemporâneas que necessitam de uma infraestrutura potente que permita o fluxo rápido e eficiente dos dados em constante transmissão” (FERREIRA, 2018, p. 22). Para Andronico *et al*. (2011), a camada mais importante da infraestrutura eletrônica é a rede de colaboração humana entre comunidades científicas de pesquisadores que trabalham juntos em problemas multidisciplinares complexos sem precedentes cujas soluções são altamente benéficas para a sociedade e o progresso em geral.

As dimensões devem ocorrer de maneira inter-relacionada e interdependente, como orientam as setas indicadas na pirâmide, para que a proposta de gerar novas práticas de produção do conhecimento voltadas para a pesquisa colaborativa seja efetivada.

Neste estudo, partiu-se do entendimento que a infraestrutura necessária a ser observada na construção de diretrizes para estruturação e publicação de dados de pesquisa científica, registrados em cadernos de pesquisa, é composta por um conjunto de tecnologias semânticas de uso distribuído para apoiar o trabalho colaborativo em redes entre cientistas na produção e uso intensivo de dados. Em consonância com o modelo de Andronico *et al*. (2011), as dimensões discutidas neste estudo são: Dados de Pesquisa Científica, Tecnologia e Colaboração Científica em Rede.

3.2.1 Dimensão Dados de Pesquisa Científica

A dimensão Dados de Pesquisa Científica corresponde ao grande volume de dados produzidos por novas tecnologias de alto rendimento. Segundo Hey e Trefethen (2003), esse dilúvio eminente de dados é o fator essencial para o desenvolvimento da *e-Science*.

Os dados constituem a base para o desenvolvimento da pesquisa científica, a qual visa produzir conhecimentos contribuindo para o avanço da ciência e para o desenvolvimento social. Além disso, os dados de pesquisa assumem um papel importante enquanto fonte de informação atualizada e contribui de forma decisiva no desenvolvimento da ciência e tecnologia. Contudo, Bell (2009) relata que os dados nos quais as teorias científicas se baseavam eram frequentemente deixados em cadernos individuais ou armazenados em mídias magnéticas que, eventualmente, se tornam ilegíveis e inacessíveis à comunidade científica.

No entanto, os dados que eram tidos como rascunhos após suas análises e publicação dos resultados em seu formato final passam a ser garantia de verificação dos resultados de uma investigação, além de uma poderosa fonte de informação para novas descobertas científicas. Nessa direção, Silva (2019) ressalta que os pesquisadores, no transcurso do seu trabalho, buscavam por documentos em fontes como repositórios e bases de dados, hoje

requerem também os dados das pesquisas. Para Silva (2019, p. 2), "o acesso on-line permite apresentar os resultados de uma pesquisa de maneira mais ampla e completa, o que representa um enorme potencial para o avanço científico".

Na *e-Science* novos dispositivos experimentais de alto rendimento cada vez mais estão sendo implantados em muitas áreas da ciência, o que levará a um verdadeiro dilúvio de dados científicos (HEY; HEY, 2006). Nesse cenário, gerenciar tais vastos conjuntos de dados se torna fundamental para a promoção de atividades científicas mais abertas e transparentes à comunidade científica em geral. Esse gerenciamento envolve ações, boas práticas, padrões e tecnologias internacionalmente reconhecidas para sustentar a ciência intensa em dados de pesquisa científica e colaborativa em rede, facilitando tanto a recopilação das pesquisas científicas quanto a aplicação de dados antigos em um novo contexto.

Nesse sentido, Jim Gray – durante a palestra proferida em 11 de janeiro de 2007, na NRC-CSTB[3] - mencionou ao menos três ações de gestão consideradas essenciais para o uso de dados: captura, curadoria e análise. A captura refere-se à coleta de dados gerados por instrumentos, redes de sensores ou simuladores de dados, e armazenados em bancos de dados. Nesse processo, os cientistas necessitam de apoio de mecanismos de pesquisa especializadas e poderosas ferramentas para mineração de dados. Para Araújo Júnior (2007, p. 58), "a mineração de dados pode ser vista como o processo de descoberta de novas correlações, padrões e tendências significativas por meio de análise minuciosa de grandes conjuntos de dados".

Após a captura, os dados precisam passar pelo processo de curadoria, o qual consiste no "gerenciamento de dados científicos primários, permitindo que, a partir da preservação e manutenção dos registros existentes em um centro, seja possível agregar valor aos dados científicos e permitir seu compartilhamento e reuso" (MEDEIROS; CAREGNATO, 2012, p. 319). Segundo Hey e Hey (2006), os registros são constituídos a partir da descrição de metadados relevantes, fornecendo informações sobre a proveniência, o conteúdo e as condições em que os dados foram produzidos. Bell (2009) reforça que sem os metadados definidamente anotados, a interpretação é implícita e é possível que os dados sejam perdidos. Silva (2019, p. xi) complementa que "um dos componentes da garantia de integridade dos registros científicos são os metadados de preservação, neles deverão estar contidas as informações que darão suporte ao processo de permanência dos registros". Diante da

[3] Computer Science and Telecommunications Board (CSTB), traduzido para o Português Conselho de Ciência da Computação e Telecomunicações é o local em que o país busca avaliações independentes e informadas de computação, comunicações e politicas públicas. Disponível em: https://sites.nationalacademies.org/CSTB/index.htm

importância do elemento metadado dentro do processo de gestão de dados, Bell (2009) enfatiza que os esquemas e os metadados são necessários para a integração entre instrumentos, experimentos e laboratórios.

Para Bell (2009) uma das preocupações da *e-Science* é a proveniência de dados em longo prazo. Nessa perspectiva, Sayão e Sales (2015) recomendam que ao utilizar dados já existentes seja necessário registrar as informações sobre as suas origens, bem como a relação entre esses dados e os dados que estão sendo coletados na pesquisa em curso. Os autores alertam que se a coleção de dados for combinada com dados já existentes, cabe definir como será assegurada a compatibilidade de formatos. Os formatos de arquivo, nesse contexto, referem-se a formatos legíveis por programas de computador. Isto porque toda informação digital é planejada para ser interpretada por um programa de computador para que seja compreendida e existente. Sendo assim, "os dados digitais são ameaçados pela obsolescência tecnológica do ambiente de hardware e de software necessários à interpretação deles" (SAYÃO; SALES, 2015, p. 59). Nesse sentido, os autores enfatizam que, para que haja a compatibilidade e o acesso de longo prazo, a recomendação é converter os dados para formatos padronizados, além de considerar o uso de formatos não proprietários, abertos e amplamente utilizados pela comunidade científica. O uso desses elementos e orientações permite que os dados sejam interpretados por vários programas, além de possibilitar o intercâmbio, a preservação e a proveniência de longo prazo dos dados (SAYÃO; SALES, 2015).

A terceira ação mencionada por Gray (2009) é a análise dos dados, a qual envolve, segundo Appel (2014, p. 15), "o uso de tecnologias computacionais distribuídas e compartilhadas, os *grids*, as quais permitem o uso de recursos tecnológicos ou habilidades analíticas de cientistas posicionados em diferentes localidades".

Para os pesquisadores, a gestão adequada dos dados científicos permite gerar novos campos de pesquisa. Na mesma direção, as novas pesquisas geram dados que necessitam ser organizados e entendidos. Portanto, a gestão dos dados requer um processo contínuo que identifique as dimensões, as categorias, as tendências, os padrões e as relações, revelando seu significado. "Esse processo é complexo e implica um trabalho de redução, organização e interpretação de dados, que inicia previamente na fase exploratória e que continua durante todo o ciclo da pesquisa" (SILVA, 2019, p. 3).

3.2.2 Dimensão Tecnológica

A dimensão tecnológica refere-se à construção de uma infraestrutura de computação distribuída baseada nas tecnologias como *Grids*, *Midleaware*, *Workbenches*, *Web Services*, *Virtual Research Enviroments* (VRE), tecnologias de notação e armazenamento de dados concebidos em padrões como XML.

Para Hey e Trefethen (2002), a tecnologia *Grid* é a infraestrutura necessária para viabilizar projetos *e-Science*, permitindo o compartilhamento rotineiro de recursos computacionais e de dados distribuídos e heterogêneos, além de apoiar a colaboração eficaz entre grupos de cientistas. Vaz (2011, p. 21) entende que a infraestrutura *Grid* é "uma nova geração de serviços de informação composta de *middleware*, software e hardware para acessar, processar, comunicar e armazenar grandes volumes de dados".

O *Middleaware* corresponde ao software que se encontra entre o sistema operacional e os aplicativos nele executado (MICROSOFT AZURE, 2020). Appel (2014, p. 13) explica que este software fica "disponível entre a infraestrutura computacional dos *grids*, distribuída em rede, e as aplicações de uso individual por cientistas e permite que cada usuário possa compartilhar seus recursos ou ter acesso aos recursos de outros usuários". Para Hey e Hey (2006) o *middleware grid* é responsável por permitir que os pesquisadores configurem facilmente suas próprias organizações virtuais seguras, vinculando sites de pesquisa com os quais desejam compartilhar uma variedade de recursos com acesso autenticado controlado.

A tecnologia *Web services* é um sistema de software projetado para oferecer suporte à interações interoperáveis máquina-a-máquina através de uma rede (W3C, 2014), com uma interface descrita em formato processável por máquina. Albagli, Appel e Maciel (2014) explicam que essa interface é uma camada intermediária que permite a comunicação entre o software ou sistemas presentes nos *grids*. Segundo a W3C (2014) outros sistemas interagem com o *Web service* da maneira prescrita por sua descrição, usando mensagens de protocolo simples de acesso a objetos (SOAP), normalmente transmitidas, valendo-se do protocolo de transferência de hipertexto (HTTP) com uma serialização de linguagem de marcação extensível (XML) em conjunto com outros padrões relacionados à Web.

Outra tecnologia considerada de apoio ao desenvolvimento de projetos *e-Science* é o *Virtual Research Enviroments* (VRE), em português Ambiente Virtual de Pesquisa (AVP), o qual é a infraestrutura que permite o compartilhamento, a colaboração e o reuso de dados científicos. Segundo Vaz (2011), o VRE pode ser visto como um *framework* em que ferramentas, serviços e recursos podem ser conectados. São exemplos de recursos VRE: sites

de informações gerais, *blogs*, *wikis*, fóruns de discussão, recursos para dispositivos móveis e outros. Alguns desses recursos VRE oferecem suporte à hospedagem de documentos e ferramentas de análise de dados, visualização ou gerenciamento de simulações. Vaz (2011) alerta que para um VRE ser efetivo é necessário estar integrado às políticas e à infraestrutura de pesquisa existente.

Vale ressaltar entre as tecnologias dessa dimensão, os repositórios de dados criados para a função de armazenar e difundir dados de pesquisa. Segundo Coneglian *et al*. (2018, p. 1), "os repositórios de dados são instrumentos para a disponibilização dos dados gerados no processo de desenvolvimento da pesquisa, visando armazenar, acessar, (re)usar e compartilhar os diversos recursos produzidos e em diferentes formatos". Para Silva (2019, p. 89), "uma rede de repositórios gera conexões entre comunidades, pois cada vez mais a inter-relação entre fontes de dados provenientes de distintas disciplinas encontra-se em repositórios específicos ou multidisciplinares". Nesse contexto, tem-se a ferramenta *Registry of Research Data Repositories* (Re3Data)[4] que fornece buscas de repositórios de dados. Atualmente há diversos repositórios de dados disponíveis e abertos, como Figshare[5], Dryad[6] e Zenodo[7]. Entretanto, muitas universidades, institutos e grupos de pesquisa estão iniciando seus próprios repositórios de dados de pesquisa como meio para o armazenamento, preservação e acesso. Para esses repositórios que estão sendo criados, recomenda-se o uso de tecnologias abertas e amplamente reconhecidas para que facilite a interoperabilidade entre sistemas.

Para Oliveira e Silva (2016), na *e-Science*, os pressupostos para uma ciberinfraestrutura[8] tecnológica sustentável exigem a adoção de modelos e tecnologias abertas. Entre as tecnologias e normas indicadas pelas autoras para a execução de cada estágio do ciclo de vida dos dados científicos, encontram-se o agrupamento de plataformas abertas, padrões de metadados para descrição dos dados, protocolos de interoperabilidade para a integração dos metadados, identificadores digitais para assegurar a qualidade dos dados, linguagens de marcação HTML e XML para favorecer a vinculação de dados; normas que proporcionam a descrição de qualidade e recursos automáticos para citação, além de mecanismos de atribuição de licenças públicas.

Essas tecnologias e metodologias compõem o conjunto de elementos fundamentais e obrigatórios no desenvolvimento de ditretrizes semânticas para publicação de dados

[4] Re3Data. Disponível em: http://re3data.org/
[5] Repositório de dados Figshare. Disponível em: https://figshare.com/
[6] Repositório de dados Dryad. Disponível em: https://datadryad.org/stash
[7] Repositório de dados Zenodo. Disponível em: https://zenodo.org/
[8] Versão de infraestrutura eletrônica adotada nos Estados Unidos.

científicos anotados em cadernos de pesquisa, as quais serão descritas nos capítulos 4 - dados de pesquisa científica e 6 - Web Semântica, respectivamente.

3.2.3 Dimensão Colaboração Científica em Rede

As novas práticas do fazer científico referem-se às técnicas colaborativas entre pesquisadores de diferentes áreas com propósito de construção do conhecimento, independente da localização geográfica dos participantes. Para Sonnenwald (2007, p. 645, tradução nossa), a colaboração científica trata-se da "interação que se situa em um contexto social entre dois ou mais cientistas, a qual facilita o compartilhamento do propósito e da ação de execução de tarefas, respeitando a um objetivo superordenado e compartilhado mutuamente". Entende-se que o fazer científico de forma colaborativa possibilita a democratização do saber, a otimização de esforços e a multiplicação de resultados.

O trabalho colaborativo no meio científico não é uma característica exclusiva do quarto paradigma da ciência, pois Ferreira (2018, p. 59) reforça a colocação de Meadows (1999, p. 107) ao declarar que "apesar da existência de pesquisadores solitários, nos primórdios da ciência houve a colaboração desde o princípio". Na sequência, a autora destaca a afirmação de González Alcaide *et al*. (2013, p. 13) que "a colaboração científica não é um fenômeno recente, os cientistas têm trabalhado cooperativamente desde que existe a ciência". Segundo Sonnenwald (2007), o termo 'colaboração científica' originalmente designava um 'laboratório sem paredes' permitindo que os cientistas realizassem pesquisas mesmo estando em regiões geograficamente dispersas. Inicialmente os pesquisadores forneciam acesso remoto a instrumentos científicos. Ao longo do tempo, a visão dos colaboradores se expandiu e a definição passou a basear-se em rede. Nesse sentido, pode-se compreender que a colaboração científica em rede corresponde a uma conexão de pessoas de diferentes áreas do conhecimento cooperando para o desenvolvimento de uma pesquisa comum. Para Sonnenwald (2007), a colaboração em rede promove o contato entre pesquisadores que são conhecidos e desconhecidos um do outro e fornece acesso a fonte de dados, artefatos e ferramentas necessárias para realizar tarefas de pesquisa.

Os autores Meadows (1999), Sonnenwald (2007) e Ferreira (2018) defendem que os estudos realizados em conjunto apresentam mais qualidade, confiança e visibilidade à pesquisa científica. Para Meadows (1999), a literatura construída em colaboração oferece mais qualidade à pesquisa científica quando comparada àquelas realizadas por pesquisadores que trabalham isoladamente, pois

> as razões básicas do trabalho em equipe encontram-se no crescimento e especialização da pesquisa, a qual demanda uma gama de conhecimentos e o acesso a recursos consideráveis (em termos de pessoal e finanças) que se situam além das possiblidades de uma única pessoa (MEADOWS, 1999, p. 109).

Segundo Sonnenwald (2007) a colaboração científica tem o potencial de resolver problemas complexos, ampliar o escopo de um projeto e promover a inovação. Isto porque, conhecimentos adicionais são disponibilizados. Além disso, o trabalho feito em colaboração pode aumentar a confiabilidade científica e a probabilidade de sucesso, pois mais de uma pessoa está atestando a validade, a precisão e a qualidade da pesquisa. Nessa direção, Ferreira (2018, p. 61) garante que "a qualidade na pesquisa e visibilidade na comunidade científica são alguns dos ganhos advindos do trabalho em colaboração".

No contexto da *e-Science*, as novas formas de colaboração científica ocorrem a partir do desenvolvimento e difusão das tecnologias da informação e comunicação, especialmente das plataformas digitais, as quais permitem a formação de comunidades virtuais (ALBAGLI; APPEL; MACIEL, 2014). Para Andronico *et al*. (2011, p. 155, tradução nossa),

> Uma comunidade virtual de pesquisa é um grupo amplamente disperso de pesquisadores e instrumentos científicos associados que trabalham juntos em um ambiente virtual comum. Esse novo tipo de ambiente científico, geralmente chamado de "colaborativo", baseia-se na disponibilidade de redes de alta velocidade e acesso à banda larga, ferramentas virtuais avançadas e tecnologias *Grid-middleware* que, no total, são os elementos das e-Infraestruturas.

Nesse cenário, Ferreira (2018, p. 58) expõe que "a colaboração pode ocorrer em massa, quando são utilizados ambientes colaborativos digitais e são oferecidos aos usuários plataformas digitais para criar, editar e compartilhar conteúdos próprios".

Do ponto de vista da formação de redes colaborativas em ciência, as quais se valem intensivamente da Internet, Dutton (2008) classificou três tipos de ações que se concentram no suporte à colaboração por meio de:

a) Compartilhamento - a capacidade de criar documentos e objetos vinculados em uma rede distribuída, reconfigurando como e quais informações são compartilhadas com quem.
b) Contribuição - a capacidade de empregar aplicativos de rede social da Web para facilitar a comunicação em grupo, remodelando assim quem contribui com informações para o grupo coletivo.

c) Co-criação - implica na habilidade de colaborar por meio de redes para facilitar o trabalho cooperativo em direção a objetivos comuns, desse modo reconfigurando o sequenciamento, composição e definição de papéis dos colaboradores.

Essas ações podem ser usadas a partir do (a) compartilhando de documentos de hipertexto, dados e outros objetos digitais; (b) implantação de ferramentas de rede social para apoiar a colaboração e gerar conteúdo do usuário; e (c) aplicação de software colaborativo para apoiar a co-criação cooperativa (DUTTON, 2008).

No que se refere aos tipos de colaboração científica remota com base nas novas TICs foram mencionadas por Sonnenwald (2007): as comunidades virtuais de prática, comunidades virtuais de aprendizagem, sistemas de instrumentos compartilhados, sistemas de dados comunitários, sistemas abertos de contribuição comunitária, centros de pesquisa distribuídos e projetos de infraestrutura comunitária. Entre os exemplos de aplicativos de TICs usados para oferecer suporte à colaboração incluem: e-mail, mensagens instantâneas, *wikis*, *blogs*, videoconferência, cadernos eletrônicos de laboratório, acesso remoto compartilhado a instrumentação, ferramentas de agendamento para organizar experimentos de laboratórios.

A título de exemplo, a Wikipédia, enciclopédia on-line global, é uma ferramenta de colaboração aberta, na qual diversas pessoas colaboram para o enriquecimento do assunto e geração do conhecimento. Para Mendes (2009), vários cientistas estão trabalhando em projetos *Wiki* para gerar livros criados por voluntários, da mesma forma, vários cientistas trabalham em projetos científicos mundo a fora para a descoberta da cura de várias doenças. Nota-se que o poder computacional, as capacidades da rede e a disponibilidade crescente da Web criaram uma explosão sem precedentes de participação e colaboração global (MENDES, 2009).

Para Hey e Hey (2006), além de poder acessar informações de diferentes locais, os cientistas agora querem poder usar recursos de computação remota, integrar, federar e analisar informações de muitos recursos de dados diferentes e distribuídos, além de acessar e controlar equipamentos experimentais remotos. Os autores expõem que a capacidade de acessar, mover, manipular e extrair dados é o requisito central desses novos aplicativos de ciência colaborativa, sejam eles mantidos em banco de dados simples ou gerados por aceleradores ou telescópios ou dados coletados em tempo real de redes de sensores potencialmente móveis.

As plataformas de tecnologias distribuídas e os aplicativos de suporte à pesquisa potencializam a criação de conhecimento e novas descobertas, no entanto, Albagli, Appel e Maciel (2014) esclarecem que a adoção de plataformas computacionais na colaboração científica não se reduz ao componente tecnológico. Os autores enfatizam a importância aos

novos usos, os quais implicam transformações nos métodos e estruturas lógicas da pesquisa e, logo, em seus resultados, em um processo de aprendizado e inovações contínuas.

A dimensão colaboração em rede ocorre de maneira inter-relacionada com as dimensões tecnológica e dados de pesquisa científica. Após a descrição das dimensões – tecnologia, dados de pesquisa científica e colaboração científica em rede – foi possível identificar os elementos conceituais e práticos presentes nas dimensões da *e-Science* que influenciam no comportamento das escolhas de formatos, padrões e diretrizes para a publicação de dados de pesquisa científica anotados em cadernos de laboratório.

Dessa forma, passa-se ao estudo dos Dados de Pesquisa Científica nas suas diferentes tipologias e abordagens enquanto componente central desta pesquisa.

4 DADOS DE PESQUISA CIENTÍFICA

Este capítulo discorre sobre os Dados de Pesquisa Científica gerados a partir de instrumentos computacionais, experimentais, observacionais, dentre outros que sejam sem análise e interpretação do autor que os gerou. Parte-se do pressuposto de que os dados de pesquisa científica quando bem gerenciados em todos os estágios de seu ciclo de vida e publicados a partir das recomendações e tecnologias da Web Semântica e *Linked Open Data*, possibilitam a recuperação, o acesso, o compartilhamento e o reuso de dados, favorecendo novas descobertas. Sendo assim, buscou-se embasamento teórico sobre a temática dados de pesquisa e subdividiu o capítulo em: 4.1 aspectos conceituais e tipológicos; 4.2 dados abertos; 4.3 ciclo de vida dos dados; 4.3 princípios e diretrizes para publicação de dados de pesquisa.

Na seção 4.1 os dados de pesquisa científica são apresentados em seus aspectos conceituais e tipológicos, os quais são classificados segundo os seus propósitos de idealização, origem, natureza e grau de estruturação.

Na seção 4.2 são apresentadas diretrizes que fundamentam a abertura dos dados, bem como as licenças, formatos e linguagens que caracterizam sua abertura legal e técnica.

A seção 4.3 aborda o ciclo de vida dos dados sob as novas práticas científicas e da importância do gerenciamento para o compartilhamento efetivo de dados de pesquisa. Na sequência, apresenta abordagens de diferentes modelos de ciclo de vida, com ênfase nos modelos *Curation Lifecycle Model* (*DCCCuration*) e o *DataONE Data Lifecycle* como instrumento de apoio ao gerenciamento.

A seção 4.4 trata das diretrizes com as orientações para o adequado gerenciamento e publicação de dados de pesquisa na Web, com ênfase nos Princípios FAIR e Melhores Práticas para Publicação de Dados na Web, recomendadas pelo W3C.

4.1 ASPECTOS CONCEITUAIS E TIPOLOGICOS

O que alicerça as publicações científicas refere-se a dados sem inferências, discussões e interpretações gerados durante o decurso de uma pesquisa. Esses dados apresentam-se na literatura com diferentes terminologias, como: dados de pesquisa, primários, brutos, e de investigação. Na literatura brasileira destacam-se estudos com a terminologia ‘dados de pesquisa’ realizados por Sales (2014), Sayão e Sales (2015), Silva, Santarem Segundo e Silva (2018) e estudos com a terminologia ‘dados científicos’ realizados pelos autores Costa (2017), Silva (2019), Dias e Oliveira (2019), Silva Júnior e Santos (2019), Córdula e Araújo (2019). Partindo da compreensão conceitual de que o termo dados de pesquisa pode estar relacionado com pesquisas de cunho opinativo e dados científicos remeterem a dados cientificamente comprovados, neste estudo nomeou-se ‘Dados de Pesquisa Científica’ para se referir a dados que ainda estão sendo pesquisados e são de natureza científica. Entretanto, a terminologia ‘dados de pesquisa’ poderá ser adotada no decorrer deste estudo em algumas frases para facilitar a leitura, porém o entendimento é amplo e de natureza científica. Nesse contexto, Borgman (2010) ressalta que a noção de dados de pesquisa também pode variar consideravelmente entre os colaboradores, e ainda mais entre as áreas do conhecimento e propósitos pelos quais foram criados.

Nesse sentido, a Organization for Economic Co-Operation and Development (OECD) (2007, p. 13, tradução nossa) define dados de pesquisa como “registros factuais usados como fonte primária para a pesquisa científica e que são aceitos pelos pesquisadores como necessários para validar os resultados do trabalho cientifico”. Nessa mesma direção, Torres-Salínas, Robinson-García e Cabezas-Clavijo (2012, p. 175) reforçam que “os dados primários são todos aqueles materiais registrados durante uma pesquisa, reconhecidos pela comunidade científica e que servem para certificar os resultados alcançados”.

Segundo o National Research Council (NRC) (1999, p. 15, tradução nossa), “os dados podem ser constituídos de fatos, números, letras e símbolos que descrevem um objeto, ideia, condição, situação ou outros fatores”. Para Silva (2019, p. 22) “podem ser de tipos numéricos, descritivo ou visual e reproduzir-se em formato de papel (incluindo notas de pesquisa em cadernos, fotografias etc.) ou digital”. O autor descreve características dos dados segundo o propósito de idealização, sendo: representativos, quando são gerados, por exemplo, para medições de variáveis, tais como a idade, a altura, o peso, a cor, a opinião etc.; implicados ou derivados, quando são produzidos a partir de outros dados, como na mudança porcentual no

tempo calculado mediante a comparação dos dados de dois períodos; quando há possibilidades de armazenamento tanto em formato analógico ou digital.

Para Borgman (2010, p. 3, tradução nossa), “alguns tipos de dados têm valor imediato e são duráveis, alguns adquirem valor ao longo do tempo, alguns têm valor transitório e outros são mais fáceis de serem recriados do que curados”. Torres-Salínas, Robinson-García e Cabezas-Clavijo (2012, p. 175) especificam que há dados “que devem provir de uma fonte única e devem ser difíceis de obter novamente por serem próprios de um momento ou circunstâncias irreplicáveis de uma forma exatamente igual”.

A partir da interpretação de entidades de difusão internacional - como o National Institutes of Health (ThIH) dos Estados Unidos, a National Science Board (ThSBD) e a Organization for Economic Cooperation and Development (OECD) -, Silva (2019, p. 22) assinala as principais definições para dados científicos, como:

> - O material registrado durante o processo investigador, reconhecido pela comunidade científica e que serve para certificar os resultados da pesquisa que se realiza;
> - O material que provêm de uma única fonte e é difícil, ou impossível, obtê-lo novamente;
> - Aquele que pode admitir muitas formas (textos, números, imagens fixas ou em movimento, dentre outras) com atributos ou características que descrevem pesquisas e entidades. (SILVA, 2019, p. 22).

Muitas das variações observadas dependem da categoria dos dados. A National Science Board (NSB) (2005) classificou os dados segundo a sua origem em observacional, computacional, experimental e registros.

> Os dados observacionais incluem medidas meteorológicas e levantamentos de altitude, que podem ser associados a lugares e horários específicos ou podem envolver múltiplos lugares e tempos. Os dados computacionais resultam da execução de um modelo de computador ou simulação, seja física ou realidade virtual cultural. Os dados experimentais incluem testes laboratoriais. Os registros do governo, negócios e vida pública e privada também fornecem dados úteis para pesquisa científica, social científica e humanística (NATIONAL SCIENCE BOARD, 2005, p. 19, tradução nossa).

Para Green, MacDonald e Rice (2009), os dados podem ser classificados, segundo a sua natureza, em: números, imagens, vídeos, software, algoritmos, equações, animações ou modelos e simulações. Segundo a sua fase de pesquisa, em: dados brutos ou preliminares (*Raw data*), os quais vêm diretamente dos instrumentos científicos; dados derivados, que são resultados do processo ou combinação de dados brutos ou de outros dados; e dados canônicos ou referenciais, que são coleções de dados consolidados e arquivados geralmente em grandes centros de dados, por exemplo, sequência genética, estrutura química, etc.

Os dados são classificados ainda segundo o grau de estruturação. Para Rautenberg, Souza, Dall'Agnol e Michelon (2018) os dados estruturados seguem um formato de armazenagem identificável como os disponibilizados em tabelas contendo linhas e colunas, facilitando seu processamento e sua recuperação por ferramentas computacionais. Para Silva (2019, p.31) esta categoria de dados "segue um esquema com estrutura determinada, definida por tabelas que representam associações entre um termo e sua relação hierárquica". Para o autor, "os dados semiestruturados incluem imagens, textos, documentos de texto e outros objetos que não formam parte de uma base de dados. Não há um modelo de dados, um esquema predefinido e, portanto, não é possível manter uma estrutura de base relacional" (SILVA, 2019, p.32). Enquanto que os dados não estruturados não possuem uma formatação específica, o que dificulta o seu processamento. Entre os exemplos de dados não estruturados incluem as mensagens de e-mails ou *blogs*, imagens e documentos de texto (RAUTENBERG; SOUZA; DALL'AGNOL; MICHELON, 2018).

4.2 DADOS ABERTOS

Os dados produzidos como parte de pesquisas científicas para serem considerados abertos seguem diretrizes internacionais em suas publicações, pois segundo a Open Defitinion (2004, on-line),

> os dados são considerados abertos quando podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição à fonte original e compartilhamento pelas mesmas licenças em que as informações foram apresentadas.

Nessa perspectiva, a Open Knowledge Foundation (OKF) (2004, on-line) apresenta diretrizes fundamentais para que os dados sejam considerados abertos:

> - **Disponibilidade e acesso:** os dados devem estar disponíveis como um todo e a um custo razoável de reprodução, de preferência através de download pela Internet. Os dados também devem estar disponíveis de forma conveniente e modificável.
> - **Reutilização e redistribuição:** os dados devem ser fornecidos sob termos que permitam a reutilização e redistribuição, incluindo a mistura com outros conjuntos de dados. Os dados devem ser legíveis por máquina.
> - **Participação universal:** todos devem ser capazes de usar, reutilizar e redistribuir - não deve haver discriminação contra campos de atuação ou contra pessoas ou grupos. Por exemplo, restrições "não comerciais" que impediriam o uso "comercial" ou restrições de uso para determinados fins (por exemplo, apenas na educação) não são permitidas.

Para atender a essas diretrizes, os dados precisam ser publicados e disponibilizados sob licença de dados abertos como as *Creative Commons* (CC), do português Licença Comum

Criativa e as *Open Data Commons* (ODC), do português Dados Abertos Comuns para explicitar as permissões de uso. Para Silva (2019, p. 33) "as *Creative Commons* têm como propósito motivar aos criadores para que definam os termos nos quais se podem usar suas obras, com que direitos e sob quais condições". Rautenberg, Souza, Dall'Agnol e Michelon (2018) explicam que as *Creative Commons* permitem cópia, reuso, distribuição e modificação do dado original. Para Silva (2019) as *Open Data Commons* compreendem uma série de licenças para ajudar a gerar e usar dados abertos.

A Open Definition (2004) cita as seguintes licenças em conformidade com os princípios estabelecidos na definição de dados abertos:

Quadro 07 - Licenças *Creative Commons* e *Open Data Commons*

Tipo de Licença	**Creative Commons License**	**Open Data Commons License**	**Descrição da Licença**
Domínio Público	*Creative Commons CCZero* (CC0)	*Open Data Commons Public Domanin Dedication and Licence* (ODC PDDL)	Pretende ser uma "dedicação de domínio público", ou seja, uma renúncia a **todos os** direitos, incluindo os de atribuição.
Atribuição	*Creative Commons Atribution* (CC-BY)	*Open Data Commons Attribution License* (ODC-BY)	Permite a redistribuição e reutilização de uma obra licenciada, desde que o criador seja creditado adequadamente.
Atribuição, compartilhar igual (*share alike*)	*Creative Commons Atribution Share-Alike* (CC-BY-SA)	*Open Data Commons Open Database License* (ODbL)	Permite a redistribuição e a reutilização de uma obra licenciada nas condições em que o criador é devidamente creditado e que qualquer trabalho derivado é disponibilizado sob "a mesma licença, semelhante ou compatível".
Atribuição, não comercial	*Creative Commons License* (CC-BY-NC)		Permite copiar, distribuir, exibir e executar a obra e fazer trabalhos derivados dela, desde que não sejam comercializados.
Atribuição, não derivados	*Creative Commons License* (CC-BY-ND)		Licença que permite distribuição, mas não modificar o original ou alterar dados.
Atribuição, não comercial, compartilhar igual (*share alike*)	*Creative Commons License* (CC-BY-NC-SA)		Deve proporcionar a atribuição, reutilizar o conteúdo somente com fins não comerciais, e colocar uma licença para compartilhar similar em trabalhos derivados.
Atribuição, não comercial, não derivados	*Creative Commons License* (CC BY-NC-ND)		Não se pode modificar o original ou utilizá-lo comercialmente e deve proporcionar a atribuição.

Fonte: Adaptado de Open Definition (2004) e Silva (2019, p. 34).

Essas licenças caracterizam a abertura legal dos dados de pesquisa científica e permitem o seu livre acesso, reuso e distribuição.

Além da atribuição das licenças, outro aspecto que caracteriza os dados abertos é a publicação a partir de formatos e linguagens abertas e legíveis por máquinas, tais como a *Resource Description Framework* (RDF), *Extensible Markup Language* (XML), *Comma Sepated Values* (CSV), *JavaScript Object Notation* (JSON), *Open Document Spreadsheet* (ODS), *Scalable Vector Graphics* (SVG) e *Geography Markup Language* (GML). Tais formatos e linguagens garantem a abertura técnica dos dados de pesquisa científica, sendo possível o acesso por uma faixa de seleção. Segundo Rautenberg, Souza, Dall'Agnol e Michelon (2018), não se deve haver cobrança pelo uso dos dados, se existirem custos são apenas referentes a sua reprodução.

Os dados de pesquisa, nas suas diversas tipologias, vêm recebendo reconhecimento pela comunidade científica como parte essencial das boas práticas de pesquisa. De acordo com a OECD (2007) - fórum que aprovou a primeira publicação oficial sobre 'Princípios e diretrizes para acesso a dados de pesquisa a partir de financiamento público' – os dados de pesquisa aumentam os retornos do investimento público; reforçam a investigação científica aberta; incentivam a diversidade de estudos e opiniões; promovem novas áreas de trabalho e permitem a exploração de tópicos não previstos pelos investigadores iniciais. Essa valorização aos dados oferece à pesquisa científica um papel inovador e crucial na abordagem dos desafios globais da sociedade, desde os cuidados com a saúde e mudanças climáticas até a energia renovável e gestão de recursos naturais. Segundo a OECD (2007), as trocas colaborativas entre diferentes comunidades científicas podem aumentar a velocidade e a profundidade de uma pesquisa, além de garantir ampla divulgação da mesma.

Para que se beneficie dos resultados dessas trocas colaborativas é necessário que haja uma boa gestão dos dados. Dessa forma, Simionato (2017) garante que grandes agências de fomento internacionais como a National Science Foundation, National Institutes of Health, National Endowment for the Humanities, Economic and Social Research Council, Wellcome Trust e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) exigem a apresentação do Plano de Gestão dos Dados (PGD) como requisito para análise do financiamento da pesquisa. Segundo Sayão e Sales (2015, p. 15), "um PGD constitui em um documento formal que estabelece um compromisso de como os dados serão tratados durante todo o desenvolvimento do projeto, e também após a sua conclusão". A elaboração do PGD envolve as ações relacionadas ao ciclo de vida dos dados que perpassa pelas etapas de planejamento, coleta e criação dos dados, assegura a qualidade e descreve o dado a partir do

uso de metadados apropriados, preserva em arquivos e em ambientes digitais adequados, possibilita a descoberta a partir de metadados, integra com outros dados e analisa o ciclo (SAYÃO; SALES, 2015).

Dessa forma, apresentam-se nas seções seguintes as etapas do ciclo de vida dos dados de pesquisa com o propósito de conhecer cada estágio e suas recomendações e, na sequência, as diretrizes para a publicação e gerenciamento de dados de pesquisa científica.

4.3 CICLO DE VIDA DOS DADOS

O processo da pesquisa tradicional envolve a produção, a coleta, a análise e a interpretação dos dados de pesquisa científica que, em um determinado contexto, geram significados. A partir da compreensão de que os dados brutos podem ser reanalisados e novas descobertas podem ser geradas, surgem outras etapas como a preservação e a conservação para o seu compartilhamento e reuso. Nessa perspectiva, Silva (2019, p. 43) apresenta o processo científico a partir de uma combinação de duas etapas:

> 1. O processo de pesquisa, quando se efetiva, há a produção, processamento e interpretação dos dados.
> 2. O processo de preservação de dados, que oferece sustentabilidade para o desenvolvimento de novos processos de pesquisa.

No cenário da *e-Science* todas as etapas do processo científico vem passando por mudanças, pois novas alternativas para produção, coleta, processamento, armazenamento e recuperação de dados estão surgindo. Paralelo ao movimento da *e-Science* surge a preocupação de como esses dados devem ser capturados, selecionados e conservados para o acesso e uso no futuro. Nesse contexto, Sant'Ana (2016) expõe que o acesso e o uso dos dados são fatores chave de sucesso na sociedade moderna e propõe a utilização do Ciclo de Vida dos Dados (CVD) para a delimitação de fases de acesso e de uso, mantendo-os como centro do CVD. Um modelo de CVD é estruturado pelas etapas que possibilitam a reutilização para além do projeto para o qual determinado dado foi criado, pois, segundo Sayão e Sales (2015), novos projetos podem recompilar ou adicionar novos elementos a esses dados, de forma que possam ser reusados por outros pesquisadores, reiniciando um novo ciclo de pesquisa.

Para Oliveira (2016) o ciclo de vida dos dados é um instrumento adotado para conduzir e apoiar o gerenciamento e compartilhamento de dados abertos e de natureza científica. Para a autora, o bom gerenciamento que contempla os estágios de um ciclo de vida de dados de pesquisa é fundamental para se obter qualidade e excelência na pesquisa, possibilitando o compartilhamento e seu reuso no futuro dos dados.

4.3.1 Modelos de Ciclo de Vida de Dados

Os modelos de ciclo de vida dos dados apresentam etapas que facilitam o gerenciamento em todas as fases do processo científico, desde a coleta até o compartilhamento para o acesso, uso e reuso futuro. Esses modelos são elaborados de acordo com o propósito, práticas de diferentes domínios e interesses institucionais. As etapas e fluxos podem ser diferentes de acordo com cada modelo.

Apesar da variedade de modelos de ciclo de vida de dados, cada qual com seu propósito e estrutura, entre os mais recorrentes na literatura e revisados por Ball (2012) encontram-se os modelos apresentados no quadro 04. Além desses, destaca-se a nível nacional o modelo proposto por Sant'Ana (2013; 2016), o qual apresenta a necessidade da construção de uma estrutura pensada entre os usuários e suas necessidades informacionais para que haja a efetiva disponibilidade, acesso e uso.

As abordagens dos modelos são apresentadas no quadro 08.

Quadro 08 - Modelos de Ciclo de Vida de Dados

Modelos de ciclo de vida dos dados	Abordagem
Curation Lifecycle Model / *DCCCuration*	Trata de maneira específica das necessidades relacionadas com a curadoria digital (SILVA, 2019, p. 49).
I2S2 *Idealized Scientific Research Activity Lifecycle Model*	Concentra nas atividades da pesquisa, publicação e administração do pesquisador, com visão geral da atividade de arquivamento (BALL, 2012, p. 5).
DataONE *Data Lifecycle*	Serve como uma estrutura subjacente para o desenvolvimento de ferramentas, serviços e materiais educacionais. Foi projetado para ser independente de domínio (WIGGINS *et al.*, 2013).
DDI Combined Lifecycle Model	Um modelo combinado para dados de pesquisa, particularmente dados de ciências sociais (BALL, 2012, p. 7).
ANDS Data Sharing Verbs, da Australian National Data Service	Apresenta um conjunto de verbos [etapas] para projetar e estruturar serviços flexíveis de compartilhamento de dados em um ambiente heterogêneo (BURTON; TRELOAR, 2009).
Uk Data Archive Data Lifecycle / UKDA	Oferece apoio aos pesquisadores, considerando como o gerenciamento de dados se relaciona com o ciclo de vida de um projeto de pesquisa (BALL, 2012, p. 9).
Research360 Institutional Research Lifecycle	Este modelo pode ser visto como um resumo de alto nível do modelo I2S2, simplificado para se parecer com o modelo de ciclo de vida do UKDA. (BALL, 2012, p. 9).
Capability Maturity Model for Scientific Data Management	Propõe práticas a partir de um modelo de gerenciamento de dados científicos definido constituído por níveis de maturidade ou de capacidades para as organizações sem processos definidos (CROWSTON; QIN, 2012).
Sant'Ana	Delimita as fases de acesso e uso de dados, mantendo os dados científicos como centro do ciclo de vida dos dados (SANT'ANA, 2013; 2016).

Fonte: Adaptado de Ball (2012).

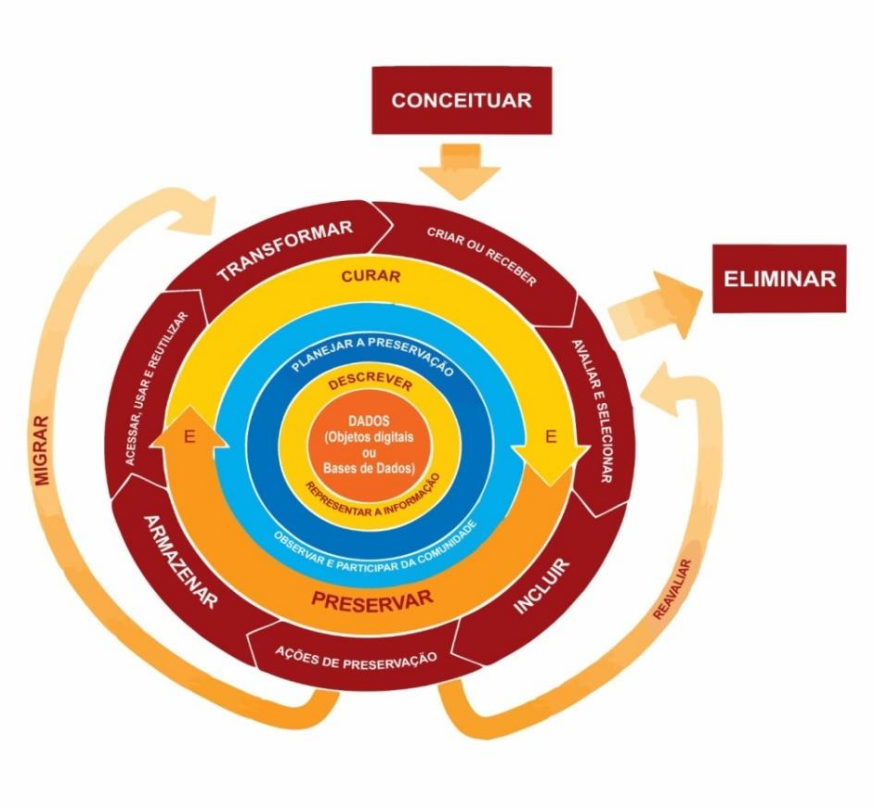
Dentre as abordagens apresentadas no quadro 08, escolheu-se os modelos *Curation Lifecycle Model* do Digital Curation Centre (DCC) e *DataONE Data Life Cycle* para descrever e exemplificar as atividades de cada etapa do ciclo de vida de dados com vista ao gerenciamento e compartilhamento de dados de pesquisa científica.

4.3.1.1 *DCC Curation Lifecycle Model*

O *DDC Curation Lifecycle Model* ou *DCCCuration* foi desenvolvido em 2007 pela Digital Curation Centre (DCC), no Reino Unido, para atender as necessidades de gerenciamento de dados especificamente no que se refere à curadoria digital. De acordo com Higgins (2008), trata-se de uma ferramenta que pode ser usada em conjunto com padrões e tecnologias relevantes para planejar atividades de curadoria e preservação de dados em diferentes níveis de granularidade. Para a autora esse modelo está sendo adotado como apoio no processo de identificação de etapas adicionais que podem ser necessárias em determinadas situações, além de garantir que os processos e políticas sejam adequadamente documentados. Ball (2012, p. 3) complementa que o modelo *DCCCuration* foi definido para "alinhar as tarefas de curadoria com as etapas do ciclo de vida de um objeto digital, destinado a ser uma ferramenta de planejamento para criadores de dados, curadores e usuários".

As etapas do modelo de ciclo de vida dos dados *DCCCuration* podem ser visualizadas na representação gráfica da figura 04.

Figura 04 - Modelo Curation Lifecycle do DCC



Fonte: Digital Curation Centre (2020, tradução nossa).

Higgins (2008) apresenta a estrutura deste modelo em três grupos de ações, sendo ações do ciclo de vida completo, ações sequenciais e ações ocasionais. Para Silva (2019) o centro do modelo destaca a importância do objeto tratado. Ainda segundo Silva (2019, p. 53), "as ações internas do ciclo de vida complementam quatro anéis concêntricos: descrição e representação da informação, planificação da conservação, observações das comunidades do processo e 'curar e preservar'". Sendo assim, esse modelo de ciclo de vida de dados para curadoria digital será apresentado a partir da exposição de Higgins (2008, p. 134-140), Ball (2012, p. 3-5) e Silva (2019, p. 49-54).

No centro do modelo encontram-se os dados, que são identificados como objetos digitais (simples ou complexos) ou bancos de dados. Os objetos digitais simples são constituídos por arquivos de texto, imagens e/ou som, juntamente aos seus identificadores e metatados relacionados. Os objetos digitais complexos são aqueles criados pela combinação de vários outros objetos digitais, como sites. As bases de dados são coleções de registros ou dados armazenados em um sistema de computador (HIGGINS, 2008, p. 137).

Para Silva (2019, p. 51) as relações entre as etapas do ciclo de vida apresentadas pelo *DDCCuration* indicam os principais níveis de ações sobre a curadoria digital dos dados,

enquanto outros modelos também contemplam as etapas de análises. As etapas de análises referem-se às ferramentas que servem de suporte para planejar todo o ciclo de vida dos dados científicos. Sendo assim, as ações seguintes (do centro para as extremidades) são observadas em três níveis de ações de ciclo de vida completo, o que graficamente constituem os quatro anéis concêntricos:

- Descrição e Representação da Informação (*Description and Representacion Information*) - descrição por meio de metadados administrativos, descritivos, técnicos, estruturais e de preservação, por meio de padrões apropriados, para garantir a descrição e controle adequado em longo prazo.
- Planejamento de preservação (*Preservation Planning*) – definição de estratégias, politicas e procedimentos para gerenciamento e administração de todas as ações do ciclo de vida de curadoria digital.
- Participação comunitária (*Community Watch & Participacion*) – acompanhamento das atividades apropriadas a comunidade e participação do desenvolvimento de padrões, ferramentas e softwares adequados.
- Curadoria e preservação (*Curate and Preserve*)– execução das ações de gerenciamento planejadas para promover a curadoria digital durante todo o ciclo de vida de dados (HIGGINS, 2008, p. 137).

Para Ball (2012) a curadoria e a preservação compõem o quarto nível do modelo, o qual descreve a maioria das ações no modelo, como também é usada para representar as ações administrativas e de gerenciamento que dão suporte à curadoria.

Ainda segundo Ball (2012), as ações sequenciais não se ocupam exclusivamente com as atividades de curadoria, mas também representam as etapas do ciclo de vida dos dados que devem ter um componente de curadoria. Iniciando com a conceituação (*Conceptualise*), a qual sinaliza para as etapas de planejamento das atividades de criação e de coleta dos dados. Para o autor, nessa fase deve-se tratar das questões de armazenado dos dados, alocação de orçamento para a curadoria e a maneira pela qual as informações importantes para a curadoria podem ser automatizadas ou simplificadas.

As etapas sequenciais que compõem o modelo iniciam com as ações de criar ou receber (*Create or Receive*), a ação de criar refere-se aos dados gerados e registrados pelos pesquisadores e receber refere-se aos dados preexistentes coletados de outras fontes. As atividades de curadoria nessa etapa concentram-se em garantir que todos os dados sejam acompanhados por metadados administrativos, descritivos, estruturais e técnicos suficientes.

Os dados recebidos devem ser conferidos e manter as mesmas políticas de coleta documentadas, dos criadores dos dados, se os padrões forem diferentes, faz-se necessário atribuir metadados apropriados.

A segunda etapa sequencial é avaliar e selecionar (*Appraise and Select*) os dados para curadoria e preservação em longo prazo. Devem-se seguir as orientações, políticas ou exigências legais documentadas. Alguns dados podem ser enviados para descarte, o que envolve a transferência dos dados para outro ambiente digital ou a eliminação definitiva, conforme orientações documentadas ou requisitos legais.

A próxima etapa é incluir (*Ingest*) os dados em um arquivo, repositório, *data center* ou outro depositário. Também, devem ser seguidas as orientações, políticas ou exigências legais documentadas. O estágio da inclusão leva ao estágio da Ação de Preservação, que envolve uma variedade de atividades, tais como controle de qualidade, catalogação, classificação, registro de metadados semânticos e estruturais, e assim por diante. Caso ocorram falhas durante a verificação do controle de qualidade, os dados são retornados ao pesquisador para uma nova avaliação. São exigências que buscam melhorias na qualidade dos dados, por exemplo, correções nos procedimentos de transferência e metadados aprimorados.

A quarta etapa sequencial é a ação de preservar (*Preservation Action*) para garantir que os dados permaneçam autênticos, confiáveis e utilizáveis, mantendo a integridade. As ações incluem limpeza de dados, validação, atribuição de metadados de preservação, representação, garantia de estruturas ou formatos de arquivo aceitáveis.

Após a conclusão do estágio da preservação, os dados passam para o armazenamento. Na etapa armazenar (*Store*), deve-se atentar para a manutenção do hardware de armazenamento, atualização da mídia e realização de cópias de *backup*. Silva (2019) alerta sobre a necessidade de conhecer as políticas do repositório para evitar danos no armazenamento de dados em longo prazo, de acordo com os padrões recomentados. Depois que os dados são armazenados com segurança, eles entram no estágio de acesso, uso e reuso.

A etapa acesso, uso e reuso (*Access, Use & Reuse*) busca garantir que os dados estejam acessíveis diariamente. Para tanto, controles de acesso e procedimentos de autenticação robustos podem ser aplicáveis. Ball (2012) chama a atenção para ações de curadoria associada a este estágio, a partir do uso da aplicação de metadados descritivos por meio de interfaces de pesquisa personalizada ou APIs públicas.

A última etapa do grupo sequencial é a transformação (*Transform*), a qual faz referência à necessidade de transformar os dados através do tempo em distintos formatos para evitar a obsolescência do software. Para Higgins (2008) a criação dos dados a partir do

original pode se dar, por exemplo, a partir da migração (*migrate*) para um formato diferente ou criar um subconjunto (*reapraise* – reavaliar), por seleção ou consulta, para criar resultados recém-derivados, talvez para publicação.

As ações ocasionais do modelo *DCCCuration* são:

- Separar ou descartar (*Dispose*) os dados que não foram selecionados para a curadoria e preservação em longo prazo, de acordo com as politicas documentadas, orientações ou requisitos legais.
- Reavaliar (*Reappraise*) os dados que falharam nos procedimentos de validação para nova avaliação e nova seleção.
- Migrar (*Migrate*) os dados para um formato diferente para garantir a imunidade à obsolescência de hardware ou software.

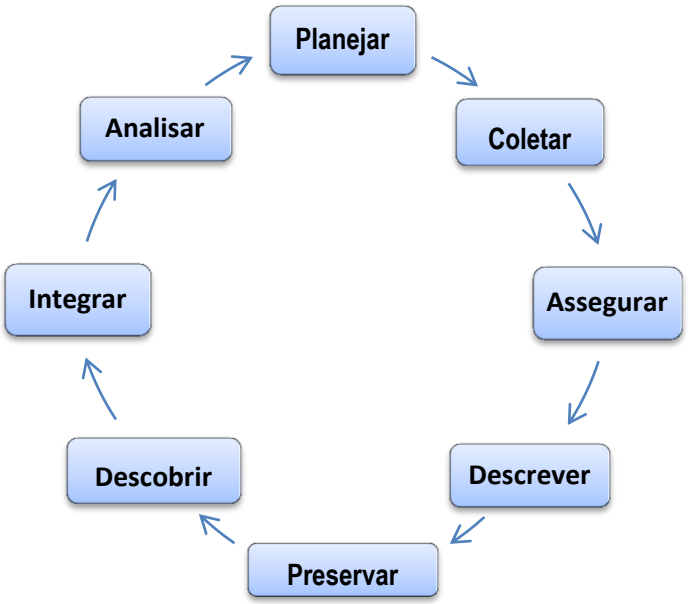
O modelo de ciclo de vida dos dados pode ser adotado em todos os seus aspectos ou partes dele, pois segundo Silva (2019) um pesquisador se dedica com frequência a todos os aspectos do ciclo de vida dos dados, inclusive pode criar novos no processo de coleta, integração, análises e sínteses dos dados já existentes. Enquanto que outros projetos podem utilizar somente uma parte do ciclo de vida, por exemplo, em um projeto que se propõe a analisar imagens do gelo da Antártida, e que poderia fazê-la durante um determinado período, outro, focado na coleta e análise dos dados primários, poderia esboçar novos descobrimentos com a integração de novos elementos.

4.3.1.2 Modelo *Data LifeCycle* do DataONE

Este modelo de ciclo de vida de dados foi construído pela Data Observation Network for Earth (DataONE) – do português Rede de Observação para a Terra - para fornecer visão ampla das etapas envolvidas no gerenciamento e preservação dos dados para uso e reuso. De acordo com Allard (2010), o DataONE é um projeto de infraestrutura cibernética que suporta o ciclo de vida completo dos dados para cientistas nos diversos domínios incorporados na ciência ambiental e ecológica. Para Allard (2010) o DataONE concentra-se em dados observacionais multidisciplinares coletados por cientistas biológicos e ambientais, redes de pesquisa nacionais e internacionais e observatórios ambientais. No entanto, a estrutura do DataONE foi projetada para ser independente de domínio, para que possa ser estendida para atender a uma gama mais ampla de domínios científicos (WIGGINS *et al.*, 2013).

O ciclo de vida de dados do DataONE é composto por oito etapas que busca contemplar todo o processo de gerenciamento de dados de pesquisa científica, conforme pode ser observado na figura 05, na sequência são apresentadas as suas descrições.

Figura 05 - Modelo Curation Lifecycle do DCC



Fonte: Strasser *et al.* (2012, tradução nossa).

A primeira etapa do ciclo é planejar (*Plan*) a descrição de como os dados serão gerenciados e disponibilizados ao longo e após a pesquisa. Nas palavras de Wiggins *et al.* (2013, p. 3) esse estágio trata-se de "um processo interativo do planejamento do projeto, durante o qual todos os aspectos do gerenciamento de dados são analisados e as decisões são tomadas para documentação e implementação em fases posteriores". Para esse estágio a recomendação é mapear os processos e recursos do projeto para o ciclo de vida dos dados, iniciando com as metas e em seguida com a definição do plano de gerenciamento.

A segunda etapa é coletar (*Collect*) os dados primários de sensores, instrumentos de laboratório e equipamentos de medição em campo e organizá-los em planilhas ou banco de dados (BALL, 2012). Para esse estágio as recomendações se baseiam principalmente na padronização do formato da coleta e do arquivamento dos dados de maneira a ser reutilizável, no uso de códigos consistentes separados por delimitadores aceitáveis, na descrição do conteúdo e organização dos dados em um arquivo.

A terceira etapa é assegurar (*Assure*) por meio de critérios e procedimentos que a qualidade e validação dos dados sejam garantidas em todos os estágios da pesquisa para o

efetivo uso e reuso em longo prazo. Segundo Wiggins *et al*. (2013), os procedimentos de controle e garantia de qualidade visam minimizar possíveis erros no acesso aos dados. Para tal, o recomendável é estabelecer um plano de controle de qualidade incluindo critérios de descrição das condições dos dados durante a coleta, a identificação dos valores (ausente e/ou questionável), a verificação daqueles que são manualmente inseridos, como do seu formato. A qualidade dos dados é, portanto, uma atividade complexa e depende de muitos fatores, por isso é essencial documentar com o máximo de detalhes o processo de qualidade (WIGGINS *et al*., 2013).

A quarta etapa é descrever (*Describe*) os dados de forma precisa e completa adotando padrões de metadados apropriados para que sejam compreensíveis por usuários humanos e máquinas, além de serem reutilizáveis de forma independente. Sendo assim, os metadados devem incluir informações que contextualizam os dados, tais como os objetivos e resultados pretendidos; quem, quando e o local da coleta; o que incluem os dados, como foram coletados e o nível de qualidade deles. A especificação de quem coletou contribui para contatar com perguntas sobre os dados quando necessário e para atribuir os créditos adequados a todos os envolvidos. Os registros detalhados dos dados garantem que eles permaneçam utilizáveis sem perda de detalhes importantes (STRASSER *et al*., 2012).

A quinta etapa compreende a preservação (*Preserve*) dos dados, cuja atividade contínua que deve ser prevista no plano de gerenciamento e ocorre em escalas de curto e longo prazo, cada qual envolvendo diferentes abordagens e decisões (WIGGINS *et al*. 2013). A preservação de curto prazo ocorre por meio de *backups* criados manualmente ou com um sistema de armazenamento automatizado como um disco rígido externo. Os *backups* são necessários para restauração de arquivos corrompidos ou perdidos e demandam técnicas de conversão, reformulação e salvamento de dados. A preservação de longo prazo visa a reutilização de dados no futuro. Dessa forma, a recomendação é enviar os dados para um ambiente digital considerado de longo prazo, como o *data center* (WIGGINS *et al*., 2013).

A sexta etapa refere-se à descoberta (*Discover*) de dados potencialmente úteis que são alocados e obtidos, juntamente aos metadados. Para Oliveira (2016, p. 81) essa atividade está relacionada "com a identificação de outros *datasets* e repositórios que possam complementar e agregar valor ao projeto". A descoberta de dados está pautada em duas etapas: "a primeira é encontrar dados existentes para análise em conjunto com outras fontes de informação; e a segunda é tornar as informações sobre os dados disponíveis para que outros possam descobrir e acessá-los" (WIGGINS *et al*., 2013, p. 11). As pesquisas encontradas na web, de modo geral, não são específicas o suficiente para encontrar dados utilizáveis, portanto o uso de

bases de dados é recomendável. Seguir recomendações apropriadas para esse estágio é indispensável para que outras pessoas possam descobrir e acessar dados, possibilitando ampliar o potencial de uso e maior visibilidade ao projeto, bem como apoiar na tomada de decisão e desenvolvimento de políticas (WIGGINS *et al*., 2013).

A sétima etapa é relativa à integração (*Integrate*) de dados de diferentes fontes, que combina-os para formar um conjunto homogêneo que possa ser facilmente analisado. Alguns cenários para a integração de dados são: 1) dados integrados de vários projetos são necessários para abordar questões complexas; 2) dados dispersos precisam ser aumentados e combinados com dados existentes para apoiar na análise; 3) dados adicionais são necessários para contextualizar, verificar e interpretar. Para Ball (2012, p. 9) esse estágio "transforma vários conjuntos de dados diferentes em uma representação comum respondendo por diferenças metodológicas e semânticas, preservando uma trilha de proveniência".

A oitava e última etapa refere-se à análise (*Analyze*) da "aplicação de modelos estatísticos e analíticos aos dados para extrair repostas significativas para as perguntas da pesquisa" (BALL, 2012, p. 9). Para essa ação existem ferramentas tecnológicas que suportam a exploração, análise e visualização. Nessa fase, é realizada a descrição da interpretação e dos resultados alcançados na pesquisa para sua posterior geração de produtos de dados derivados (WIGGINS *et al*., 2013).

Assim como o modelo *DCCCuration*, o modelo de ciclo de vida de dados do DataONE pode ser adotado em todos os seus aspectos ou parte deles, como também podem adotá-lo, mas não seguir o caminho linear descrito nesse modelo. Strasser *et al.* (2012) exemplificam que um projeto envolvendo meta-análise pode se concentrar nas etapas de descoberta, integração e análise, enquanto um projeto focado na coleta e análise de dados primários pode ignorar as etapas de descoberta e integração. Além disso, alguns cientistas podem criar novos dados no processo de descoberta, integração, análise e síntese dos dados existentes.

Para a execução do gerenciamento de dados de modo a contemplar todas as etapas do ciclo de vida dos dados de um processo científico, faz-se necessária a adoção de princípios e tecnologias para otimizar a qualidade, o acesso e usabilidade de dados. Dessa forma, a próxima seção refere-se ao estudo dos princípios que norteam a publicação de dados de pesquisa científica de forma a torná-los, dentre outros benefícios, encontráveis, acessíveis, interoperáveis e reutilizáveis.

4.4 PRINCÍPIOS E DIRETRIZES PARA PUBLICAÇÃO DE DADOS DE PESQUISA CIENTÍFICA

O cenário da enorme produção de dados de pesquisa, exploração da tecnologia em larga escala e a colaboração em rede trouxe a preocupação de como esses dados devem ser organizados e publicados para uso no futuro. Dessa forma, pesquisadores e instituições passaram a projetar princípios e tecnologias para facilitar a gestão e publicação dos dados para serem encontráveis, acessíveis, interoperáveis e reutilizáveis.

Em âmbito internacional tem-se a Organization for Economic Co-Operation and Development (OECD) que liderou o desenvolvimento de princípios e diretrizes para facilitar o acesso aos dados de pesquisa gerados com financiamento público. As discussões iniciaram em 2004, entre os governantes de 30 países integrantes da OECD, da China, África do Sul, Israel e Rússia para discutir sobre as diretrizes internacionais de acesso aos dados de pesquisa, a partir de financiamento público, e como resultado da reunião se aprovou a *Declaration on Access to Research Data from Public Funding*. Em 2007, foi apresentado o documento *Principles and Guidelines for acess to research Data from Public Fundig*, do português Princípios e Diretrizes para Acesso a Dados de Pesquisas Provenientes de Financiamento Público. Os princípios e diretrizes propostos pela OECD pretendem servir de orientação a todos os atores envolvidos - pesquisadores, instituições de pesquisa e agências de financiamento - na tentativa de melhorar o compartilhamento internacional e o acesso aos dados de pesquisa. Para Silva (2019) os pesquisadores pretendem ajudar a superar os diversos obstáculos e desafios surgidos na raiz do intercâmbio internacional de dados, com os problemas tecnológicos, de gestão institucional ou financeira, assim como as questões relacionadas ao financiamento, produção, administração e uso dos dados (OECD, 2007).

Outro movimento de destaque que coopera para a publicação aberta de dados de pesquisa é a Open Knowledge Foundation (OKF), do português Fundação do Conhecimento Aberto, criada em 2004, para promover o acesso aos conteúdos a dados abertos. Dessa fundação surgiram novos projetos, como a *Public Domain Dedication and Lincense* (PDDL), uma licença pensada para o uso de bases de dados e a Open Knowledge Brasil, também conhecida como Rede de Conhecimento Livre, que atua com diversos projetos sobre dados abertos.

No Brasil, encontra-se a Lei de Acesso Aberto, n° 12.527, de 18 de novembro de 2011, que destina a assegurar o direito fundamental de acesso à informação. No entanto, a primeira iniciativa voltada para dados abertos no campo da pesquisa científica foi o Manifesto

de Acesso Aberto a Dados da Pesquisa Brasileira para a Ciência Cidadã, lançado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em 2016, que busca promover o acesso aberto aos dados de pesquisa e apoiar movimentos para Ciência Aberta no Brasil.

Entre as iniciativas de maior destaque na literatura encontram-se as Melhores Práticas para Publicação de Dados na Web, recomendadas pelo World Wide Web Consortium (W3C) e os Princípios FAIR designados para publicar dados de pesquisa de modo a serem encontráveis, acessíveis, interoperáveis e reusáveis. Conforme estudo de Silva, Santarem Segundo e Silva (2018) essas recomendações se complementam em termos de orientações e tecnologias. As referidas recomendações serão descritas nas seções 4.4.1 – princípios FAIR e 4.4.2 – melhores práticas para publicação de dados na Web.

4.4.1 Princípios FAIR

Os princípios FAIR, um acrônimo de *Findable*, *Acessible*, *Interoperable* e *Reusable*, originaram-se em 2014, na conferência internacional *Jointly designing a data FAIRPORT* – organizada pelas Netherlands eScience Center e Lorentz Center (Leiden) - a partir de um debate entre representantes de diversas áreas do conhecimento, entre eles pesquisadores, bibliotecários, arquivistas, editores e financiadores de pesquisas, membros da *The Future of Research Communications and e-Scholarship* (FORCE11), para melhorar o ecossistema dos dados de pesquisa e funcionar como diretrizes para aumentar a sua reutilização, no âmbito da *e-Science* (FORCE11, 2014).

O resultado dessa discussão culminou-se em quatro princípios primordiais com quinze subprincípios, os quais apresentam práticas orientadoras para publicação de dados que fossem facilmente encontráveis, acessíveis, interoperáveis e reutilizáveis, por máquinas e humanos, frente a grande quantidade de informações geradas pela ciência contemporânea intensiva em dados (FORCE11, 2014).

A ideia central dos princípios FAIR está no aprimoramento da capacidade das máquinas de encontrar e usar automaticamente os dados, além de apoiar a sua reutilização por indivíduos. Sendo assim, tais princípios incorporam características que definem que os recursos, ferramentas, vocabulários e infraestrutura de dados contemporâneos devessem ser exibidos para auxiliar na descoberta e reutilização de terceiros (FORCE11, 2014). Os princípios FAIR são:

Quadro 09 - Princípios FAIR

Para ser pesquisável /encontrável: F1. Os (meta) dados são atribuídos a um identificar globalmente exclusivo e persistente; F2. Os dados precisam ser descritos com metadados enriquecidos (definidos por R1 abaixo); F3. Os metadados incluem claramente e explicitamente os identificadores dos dados que descrevem; F4. Os (meta) dados devem ser registrados ou indexados em recursos que ofereçam capacidades de busca. **Para ser acessível:** A1. (Meta) dados devem ser recuperáveis pelos seus identificadores usando protocolo de comunicação padronizado; A1.1 O protocolo deve ser aberto, gratuito e universalmente implementável; A1.2 O protocolo deve permitir um procedimentos de autenticação e autorização, quando necessário; A2. Os metadados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis. **Para ser interoperável:** I1. (Meta) dados devem ser representados por meio de uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento; I2. (Meta) dados devem utilizar vocabulários e/ou ontologias que seguem os princípios FAIR; I3. (Meta) dados devem incluir referências qualificadas para outros (meta) dados. **Para ser reutilizável:** R1. Os metadados devem ser ricamente descritos com uma pluralidade de atributos precisos e relevantes; R1.1. (Meta) dados devem ser disponibilizados com licenças de uso de dados claras e acessíveis; R1.2. (Meta) dados devem estar associados à sua proveniência; R1.3. (Meta) dados devem estar alinhados com padrões relevantes ao seu domínio. * Destaca-se que a expressão "(meta) dados" é adotada nos casos em que o princípio deve ser aplicado a metadados e aos objetos de dados.

Fonte: Adaptado de Wilkinson *et al*. (2016).

O princípio encontrável (*Findable*) apresenta orientações para se efetivar, sendo que em F1 prever que seja atribuído metadados com identificador persistente, único e global, para cada conjunto de dados. É um exemplo de identificador persistente o *Digital Object Identifier* (DOI) adotado para atribuir direitos de propriedade intelectual e para intercambiar informações sobre essas propriedades em um ambiente digital. Em F2 a recomendação prática é de enriquecer metadados com informações de outros *datasets* de modo a aumentar as possibilidades de encontrar determinados dados ou conjuntos de dados. Segundo Henning *et al.* (2019), os conjuntos de dados, quando enriquecidos o suficiente, possibilitam que os usuários encontrem-os mesmo que não possuam o seu identificador. A orientação em F3 é

associar os dados aos metadados por meio da inclusão do identificador, pois, conforme explica Henning *et al*. (2019), não há previsão de que os dados e seus metadados estejam indexados em uma mesma plataforma digital. Em F4 recomenda-se a indexação dos dados e metadados em repositórios de dados para que sejam encontráveis (WILKINSON *et al*. 2016).

Quanto aos conjuntos de dados e/ou metadados serem acessíveis (*Acessible*), a recuperação deve ocorrer pelo identificador por meio de um protocolo padrão de comunicação, conforme recomendado em A1. Nas recomendações A1.1 e A1.2 são expostas que o protocolo de comunicação usado para dar acesso deve ser aberto, gratuito, implementável por qualquer pessoa e que possa permitir um procedimento de autenticação e autorização quando se fizer necessário. Hodson *et al*. (2018) esclarecem que a acessibilidade em FAIR não implica necessariamente que os dados sejam gratuitos, mas propõe que os princípios sejam estendidos para exigir que os metadados sejam abertos e gratuitos por padrão e para incentivar que os dados sejam disponibilizados o mais aberto possível no sentido de reduzir os limites de acesso. Em A2 orienta-se que os metadados estejam disponíveis mesmo quando os dados tornarem indisponíveis. Essa orientação é praticável quando se adota estratégias de preservação de dados. Hodson *et al*. (2018) complementam que os metadados devem ser disponibilizados como dados abertos vinculados ou coletáveis por um protocolo como o *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), para que os agregadores de dados possam agrupar e reutilizar as informações.

Para que os dados sejam interoperáveis (*Interoperable*), os princípios I1, I2 e I3 recomendam a representação de conjuntos de dados e metadados por meio de linguagens formais, acessíveis, compartilhadas e amplamente aplicáveis como RDF, XML, dentre outras, bem como a utilização de vocabulários e/ou ontologias que seguem os princípios FAIR e incluem referências qualificadas a outros dados ou metadados. As recomendações desses princípios descrevem uma interoperabilidade semântica. Para Hodson *et al*. (2018, p. 12) “os dados são descritos usando padrões e vocabulários normativos e reconhecidos pela comunidade que determinam o significado preciso dos conceitos e qualidade que os dados representam”. Para os autores, esses elementos permitem que os dados sejam reconhecidos por máquinas, de modo que os valores para um conjunto de atributos possam ser examinados em uma vasta gama de conjuntos de dados, com o conhecimento de que os atributos que estão sendo medidos ou representados são realmente os mesmos.

Para aumentar as possibilidades dos dados serem reutilizáveis (*Reusable*), os princípios R1, R1.1, R1.2 e R1.3 reafirmam que os metadados sejam ricamente descritos, com o maior número de atributos possíveis que representem os dados, além de serem

disponibilizados com licenças acessíveis, bem como os dados e seus metadados estarem associados à sua proveniência e alinhados aos padrões relevantes ao seus domínios. Espera-se encontrar informações de como os dados foram criados e de seus criadores, bem como metadados com atributos completos e inequívocos. Na interpretação de Hodson *et al*. (2018, p. 12), deve-se "incluir informação sobre processos de redução ou transformação de dados que são empregados em um determinado domínio para tornar os dados mais reutilizáveis, compreensíveis ou "prontos para a ciência".

Segundo Wilkinson *et al*. (2016), os elementos dos princípios FAIR estão relacionados, mas são independentes e separáveis e podem ser implementados em qualquer combinação, de forma incremental, à medida que os provedores de dados evoluem suas estruturas no sentido de atingir um grau maior dentro do propósito dos princípios FAIR. Os autores esclarecem que estes princípios precedem as escolhas de implementação e não sugerem tecnologias específicas para implementação. Assim, esta tese sugere correlacionar os Princípios FAIR com as melhores práticas para publicação de dados na Web, as quais seguem as diretrizes do *Linked Data*, a partir das tecnologias da Web Semântica.

4.4.2 Melhores Práticas para Dados na Web – W3C

As Melhores Práticas para Dados na Web (MPs) são uma evolução da proposta de Tim Berners-Lee em fornecer uma estrutura de significados semânticos para fazer ligações entre conjuntos de dados. As diretrizes das melhores práticas se beneficiam dos conceitos e potencialidades das tecnologias projetadas pelo World Wide Web Consortium (W3C) para tornar os dados semânticos, abertos e conectados visando à interpretação por usuários humanos e máquinas.

O Consórcio W3C recomenda um conjunto de melhores práticas para publicar e consumir dados na Web. Nesse cenário, destacam-se dois atores diretamente envolvidos no processo: os provedores e os consumidores de dados. Para Lóscio, Burle, Oliveira e Calegari (2018), os provedores de dados são os atores que publicam e compartilham dados, com acesso livre ou controlado. Enquanto que os consumidores são aqueles que buscam encontrar, usar e fazer conexões entre os dados, especialmente se os dados forem precisos, atualizados e tiverem garantia de alta disponibilidade. Os consumidores podem ser eles mesmos provedores de dados. Nesse sentido, cria-se a necessidade do estabelecimento de recomendações para o gerenciamento e publicação de dados na Web para que haja um entendimento comum entre os provedores e consumidores de dados.

Assim, Lóscio, Burle e Calegari (2017) desenvolveram 35 (trinta e cinco) Melhores Práticas para Dados na Web, (DWBP, do inglês *Data on the Web Best Practices*), como recomendação do W3C para melhorar a coerência entre provedor e consumidor, incentivar e permitir a expansão continuada da Web como um meio para o intercâmbio de dados e promover a reutilização de dados de forma confiável (LÓSCIO; BURLE; OLIVEIRA; CALEGARI, 2018). Segundo Rautenberg, Souza, Dall'Agnol e Michelon (2018), as MPs são de propósito geral para serem utilizadas independentemente de domínio e de aplicação. As melhores práticas podem ser estendidas com outras diretrizes existentes, objetivando contemplar contextos mais específicos.

A aplicação das melhores práticas garante os seguintes benefícios:

1. Reuso – é o aumento das chances de reutilização dos dados por diferentes grupos de consumidores.

2. Compreensão – possibilita aos seres humanos o entendimento sobre a natureza, a estrutura, o significado e os metadados dos dados disponibilizados.

3. Interligação – permite a criação dos relacionamentos entre recursos de dados, desde os conjuntos de dados até os itens de dados nesses conjuntos.

4. Descoberta – habilita os computadores a descobrirem automaticamente um conjunto de dados e os dados aninhados no conjunto.

5. Confiança – para os consumidores de dados, certifica que o conjunto de dados é curado ao longo do tempo.

6. Acesso – facilita o acesso aos dados atualizados em diversas formas, tanto para os usuários como para os computadores.

7. Interoperabilidade – permite que os dados publicados sejam consumidos por diferentes sistemas, nos mais variados formatos.

8. Processabilidade – habilita os computadores a processar e a manipular automaticamente os dados contidos na Web (LÓSCIO; BURLE; CALEGARI, 2017, on-line).

As trinta e cinco (35) melhores práticas para publicar dados na Web são distribuídas em categorias e, para cada melhor prática, obtém-se um conjunto de benefícios, como estão apresentados no quadro 10.

Quadro 10 - Melhores Práticas e Benefícios

Categoria	Melhor Prática	Benefícios
Metadados	Fornecer metadados para usuários humanos e máquinas	Reuso, compreensão, descoberta e processabilidade
	Fornecer metadados descritivos	Reuso, compreensão e descoberta
	Fornecer metadados estruturados	Reuso, compreensão e processabilidade
Licenças	Fornecer informações de licença de dados	Reuso e confiabilidade
Proveniência de dados	Fornecer informações completas sobre as origens dos dados e quaisquer alterações feitas	Reuso, compreensão e confiabilidade
Qualidade de dados	Disponibilizar informações sobre a qualidade de dados e adequações necessárias	Reuso e confiabilidade
Versão de dados	Atribuir versão para cada conjunto de dados	Reuso e confiabilidade
	Fornecer um histórico de versão	Reuso e confiabilidade
Identificadores de dados	Usar URIs persistentes de conjunto de dados	Reuso, interligação, descoberta e interoperabilidade
	Usar URIs persistentes como identificadores dentro de conjuntos de dados	Reuso, interligação, descoberta e interoperabilidade
	Atribuir URIs a versões de conjuntos de dados	Reuso, descoberta e confiança
Formatos de dados	Usar formatos de dados padronizados	Reuso e processabilidade
	Usar representações de dados locais neutras	Reuso e compreensão
	Fornecer dados em vários formatos	Reuso e processabilidade
Vocabulários de dados	Reutilizar vocabulários, de preferência padronizados	Reuso, processabilidade, compreensão, confiança e interoperabilidade
	Escolher o nível de formalização correto	Reuso, compreensão e interoperabilidade
Acesso a dados	Permitir o acesso completo (em massa)	Reuso e acesso
	Permitir o acesso parcial ao conjunto de dados	Reuso, acesso, interligação e processabilidade
	Disponibilizar dados em vários formatos	Reuso e acesso
	Permitir o acesso em tempo real	Reuso e acesso
	Fornecer dados atualizados	Reuso e acesso
	Explicar os motivos de quando os dados não estiverem mais disponíveis	Reuso e confiabilidade
	Disponibilizar dados através de uma API	Reuso, processabilidade, interoperabilidade e acesso
	Usar padrões da Web como base de APIs	Reuso, interligação, interoperabilidade, descoberta, acesso e processabilidade
	Fornecer documentação à medida que adicionar ou alterar uma API	Reuso e confiabilidade
	Evitar quebrar alterações na sua API	Reuso e interoperabilidade
Preservação de dados	Preservar o identificador e fornecer informações sobre o recurso arquivado	Reuso e confiabilidade
	Avaliar a cobertura de um conjunto de dados antes da sua preservação	Reuso e confiabilidade
Feedback	Coletar *feedback* dos consumidores	Reuso, compreensão e confiabilidade
	Disponibilizar publicamente o *feedback*	Reuso e confiabilidade
Enriquecimento de dados	Enriquecer dados gerando novos dados	Reuso, compreensão, confiabilidade e processabilidade
	Oferecer apresentações complementares	Reuso, compreensão, acesso e confiabilidade
Republicação	Fornecer *feedback* ao publicador original	Reuso, interoperabilidade e confiabilidade
	Seguir os termos de licença	Reuso e confiabilidade
	Citar a publicação original	Reuso, descoberta e confiabilidade

Fonte: Adaptado de Lóscio, Burle e Calegari (2017).

As melhores práticas serão apresentadas a partir das categorias pelas quais estão inseridas, bem como os seus respectivos resultados esperados e os benefícios proporcionados quando são adotadas adequadamente. Para essa apresentação, adotou-se principalmente a publicação dos autores das melhores práticas, Lóscio, Burle e Calegari (2017) e Lóscio, Burle, Oliveira e Calegari (2018).

4.4.2.1 Metadados

Os metadados são fundamentais para descrever, recuperar e acessar dados digitais na Web. Segundo Taylor (2003, p. 139), o termo metadado designa "[...] a informação estruturada que descreve atributos de recursos informacionais com o propósito de identificação, descoberta e, às vezes, administração". Dessa forma, essa categoria busca fornecer metadados para usuários humanos e máquinas, metadados descritivos e metadados estruturados, conforme recomendado nas três primeiras melhores práticas.

MP1 – Fornecer metadados para usuários humanos e máquinas

Para a publicação de dados de pesquisa científica, no contexto semântico, é necessário oferecer metadados interpretáveis pelos usuários humanos e máquinas como mencionado. Segundo Silva, Santarem Segundo e Silva (2018), o uso de metadados é indicado para encontrar e acessar objetos de dados que estejam descritos suficientemente para pessoas e máquinas distinguirem um objeto de outro. Os resultados esperados é que os humanos sejam capazes de compreender os metadados e os agentes de software sejam capazes de processá-los. De acordo com as melhores práticas, os metadados possibilitam os benefícios do reuso, compreensão, descoberta e processabilidade.

MP 2 – Fornecer metadados descritivos

Segundo Alves e Santos (2013), os metadados descritivos são utilizados para descrever, identificar e representar recursos gerais de um conjunto de dados. Segundo Lóscio, Burle, Oliveira e Calegari (2018, p. 34) "os seres humanos deverão ser capazes de interpretar a natureza do conjunto de dados e suas distribuições, e agentes de software deverão ser capazes de descobrir automaticamente conjuntos de dados e distribuições". A aplicação da MP 2 oferece os benefícios para o reuso, compreensão e descoberta dos dados.

MP 3 – Fornecer metadados estruturados

Os metadados estruturados são aqueles que estão em conformidade com um padrão previsível ou pré-determinados. Segundo Lóscio, Burle, Oliveira e Calegari (2018, p. 35), “os seres humanos serão capazes de interpretar o esquema de um conjunto de dados e agentes de software serão capazes de processar automaticamente os dados das distribuições”. Os benefícios proporcionados pelos metadados estruturados são o reuso, a compreensão e a processabilidade dos dados.

4.4.2.2 Licenças de Dados

A licença de dados estabelece o nível de permissão do proprietário quanto ao compartilhamento e reutilização. Segundo Rautenberg, Dall’Agnol e Michelon (2018), para publicar dados abertos na Web, é importante que eles sejam disponibilizados para outros usuários sob um termo legal de uso ou uma licença, na qual os dados tornem-se os mais disponíveis possíveis, dentro dos limites que a lei permite.

Rautenberg, Dall’Agnol e Michelon (2018) esclarecem que é possível utilizar uma licença padrão já existente ou disponibilizar os dados através de uma licença própria. A decisão da escolha da licença ideal para os dados deve ser baseada na forma como o proprietário deseja disponibilizar seus dados, se é permitida alteração ou somente uso, se poderão ter fins comerciais ou não, e se os dados alterados serão distribuídos sob a mesma licença. A melhor prática correspondente a indicação de licença de uso de dados é a MP4.

MP 4 - Fornecer informações sobre a licença de dados

Segundo Lóscio, Burle, Oliveira e Calegari (2018, p. 35), a recomendação desta prática é que “os usuários humanos sejam capazes de compreender a licença de dados, descrevendo eventuais restrições impostas à utilização de certos dados, os agentes de software sejam capazes de detectar automaticamente a licença de dados de uma distribuição”. Para Lóscio, Burle e Calegari (2017), as licenças de um conjunto de dados podem ser especificadas nos metadados, ou fora deles, em um documento separado ao qual estão vinculadas, com atenção às recomendações do item 4.4.2.2. Os benefícios obtidos com a aplicação desta recomendação são o reuso e a confiabilidade dos dados.

4.4.2.3 Proveniência de Dados

A proveniência de dados significa informar sua origem ou sua fonte. O produtor de dados pode não ser necessariamente o editor de dados e, portanto, coletar e transmitir esses metadados correspondentes são particularmente importantes para garantir a integridade e credibilidade de dados que estão sendo compartilhados. A melhor prática ligada à proveniência de dados é a MP5.

MP 5 – Fornecer informações de proveniências de dados

Para esta prática espera-se que os humanos conheçam a origem e o histórico dos conjuntos de dados e tenham informações para julgar a veracidade e a qualidade desses conjuntos; e que os agentes de software sejam capazes de processar automaticamente informações de proveniência. O fornecimento de informações completas sobre as origens dos dados e quaisquer alterações feitas proporciona os benefícios de reuso, compreensão e confiabilidade.

4.4.2.4 Qualidade dos Dados

Para Lóscio, Burle e Calegari (2017), a qualidade dos dados envolve diferentes tipos de dimensões de qualidade, cada uma representando grupos de características relevantes para editores e consumidores. Para Lóscio, Burle e Calegari (2017), o vocabulário de qualidade de dados (DQV, do inglês *Data Quality Vocabulary*) define conceitos como medidas e métricas para avaliar a qualidade de cada dimensão, além de várias propriedades e classes adicionais adequadas para expressar a qualidade de um conjunto de dados. A recomendação para a obtenção de qualidade dos dados é requerida na MP 6.

MP 6 – Fornecer informações sobre a qualidade dos dados

Lóscio, Burle, Oliveira e Calegari (2018, p. 35) recomendam a inclusão de metadados com informações sobre qualidade de dados de modo que “os seres humanos e agentes de software sejam capazes de avaliar a qualidade e, portanto, a adequação de um conjunto de dados para a sua aplicação”. O DQV recomenda como medida de implementação que muitos atores avaliem a qualidade dos conjuntos de dados e publiquem suas anotações, certificados e opiniões sobre o conjunto deles. O editor de um conjunto de dados deve avaliar e publicar metadados que ajudem os consumidores a determinar se eles podem usar o conjunto em seu

beneficio. Além dos editores, é recomendável que agências de certificação, agregadores de dados, consumidores de dados, também façam avaliação dos conjuntos de dados.

4.4.2.5 Versionamento dos Dados

O versionamento de dados refere-se à indicação das atualizações dos conjuntos de dados. De acordo com Rautenberg, Dall'Agnol e Michelon (2018), para tratar essas mudanças, novas versões dos conjuntos de dados podem ser criadas. Mesmo para pequenas mudanças, é importante manter as diferentes versões dos conjuntos de dados para torná-los confiáveis. Os autores orientam que os diferentes tipos de atualizações de um conjunto de dados necessitam de uma abordagem informativa e consistente para o versionamento. Dessa forma, os consumidores de dados podem entender e trabalhar com essas alterações.

MP 7 – Fornecer informações das versões dos dados

A recomendação desta prática é informar as versões de cada modificação do conjunto de dados de modo que os usuários humanos e os agentes de software identifiquem facilmente qual a versão de um conjunto de dados. Os benefícios apresentados, portanto, são o reuso e a confiabilidade dos dados.

MP 8 – Fornecer informações de histórico de versões

Para Lóscio, Burle, Oliveira e Calegari (2018, p. 37) "os usuários humanos e os agentes de software devem ser capazes de entender como o conjunto de dados muda de versão para versão e como quaisquer duas versões específicas se diferem". Esta prática também proporciona os benefícios do reuso e confiabilidade dos dados.

4.4.2.6 Identificadores dos Dados

Os identificadores, em suas diferentes formas, são responsáveis pela descoberta, o uso e a citação de dados na Web. Os identificadores uniformes de recursos (URIs) são adotados para nomear coisas e os URIs HTTP são adotados para que as pessoas possam procurar esses nomes. Os autores Lóscio, Burle e Calegari (2017, on-line) enfatizam alguns pontos sobre URIs no contexto atual:

- Os URIs possuem a função exclusivamente de identificar um recurso.
- Embora os URIs não carreguem semântica, a legibilidade por humanos é útil.

- Quando não referenciado (ou não procurado), um simples URI pode oferecer o mesmo recurso em mais de um formato.
- Um URI pode redirecionar para outro.
- A desreferência a um URI aciona um programa de computador para executar em um servidor, no qual o URI age como uma chamada a uma API. O servidor pode fazer algo como retornar um arquivo estático simples ou realizar um processamento complexo.

MP 9 - Usar URIs persistentes como identificadores de conjuntos de dados

Segundo Lóscio, Burle e Calegari (2017, on-line) "a adoção de um sistema de identificação comum permite processos básicos de identificação e comparação de dados por qualquer parte interessada, de maneira confiável". Para os autores, o resultado pretendido com tal prática é que os conjuntos de dados ou informações sobre eles sejam descobertos e citados ao longo do tempo, independente é sua disponibilidade ou do formato dos dados. Dessa forma, obtêm-se os benefícios do reuso, da interligação, da descoberta e da interoperabilidade.

MP 10 - Usar URIs persistentes como identificadores dentro de conjuntos de dados

Espera-se que os mesmos identificadores sejam utilizados por meio de diversos conjuntos de dados e que os seus identificadores também possam ser referidos por outros conjuntos de dados, ou seja, conectando vários pontos de dados sobre o mesmo recurso, criando e fortalecendo o efeito rede na Web. Para Lóscio, Burle e Calegari (2017, on-line), "os dados tornam mais valiosos se eles se referirem para outros dados sobre o mesmo tema, o mesmo local, o mesmo conceito, o mesmo evento, a mesma pessoa, e assim por diante". Os itens precisam estar relacionados em toda a Web criando um espaço global de informação acessível a humanos e máquinas. Quando tais dados identificadores são URIs HTTP, estes podem ser consultados e mais dados podem ser descobertos. Os benefícios desta melhor prática são o reuso, a interligação, a descoberta e a interoperabilidade (LÓSCIO; BURLE; CALEGARI, 2017, on-line).

MP 11 - Atribuir URIs para versões de conjuntos de dados e séries

Lóscio, Burle, Oliveira e Calegari (2018, p.38) recomendam atribuir URIs a versões individuais de conjuntos de dados assim como a séries em geral, para que "seres humanos e

agentes de software sejam capazes de se referir a versões específicas de um conjunto de dados, séries de conjunto de dados, bem como a versão mais recente de um conjunto de dados". Os benefícios esperados para esta prática são o reuso, a descoberta e a confiança.

4.4.2.7 Formato de Dados

A escolha dos formatos adequados para a publicação dos dados de pesquisa na Web garante a disponibilidade dos dados ou conjuntos de dados para uso e reuso. O Consórcio W3C recomenda melhores práticas para selecionar formatos tanto em nível de arquivos quanto de campos individuais, além de incentivar a adoção de formatos amplamente utilizados e que possam ser processados de forma simples por máquinas.

MP 12 – Usar formatos de dados padronizados legíveis por máquina

O Consórcio W3C recomenda o uso de formatos legíveis por máquinas de modo a serem capazes de ler e processar dados publicados na Web e que os usuários humanos sejam capazes de usar ferramentas computacionais para manipular os dados. Os benefícios esperados desta prática são o reuso e processabilidade de dados (LÓSCIO; BURLE; OLIVEIRA; CALEGARI, 2018).

MP 13 – Usar representações de dados que sejam independentes de localidade (*locale neutral*)

Esta melhor prática orienta disponibilizar informações sobre parâmetro de localidade para que os usuários humanos e máquinas sejam capazes de interpretar o significado de caracteres que representam datas, horas, moedas e números com precisão. Sendo assim, os benefícios adquiridos serão o reuso e boa compreensão dos dados (LÓSCIO; BURLE; CALEGARI, 2017).

MP 14 – Fornecer dados em múltiplos formatos

Ao disponibilizar os dados em vários formatos, quando mais de um se adequa ao seu uso pretendido reduz custos decorrentes da transformação de dados e minimiza a possiblidade de erros no processo de transformação. São exemplos de formatos indicados pela W3C, o RDF/XML e o Turtle. Seguindo essas orientações, os dados serão reutilizáveis e processáveis (LÓSCIO; BURLE; CALEGARI, 2017).

4.4.2.8 Vocabulários de Dados

Os vocabulários definem os conceitos e relacionamentos (termos e atributos) usados para descrever e representar uma área de interesse. Os vocabulários são adotados para classificar os termos que podem ser utilizados em uma aplicação específica, caracterizar possíveis relações e definir possíveis limitações na utilização desses termos. Segundo Rautenberg, Dall'Agnol e Michelon (2018, p. 44), "os vocabulários asseguram um nível de controle, padronização, interoperabilidade dos dados. Eles facilitam a usabilidade dos conjuntos de dados". O consórcio W3C ressalta que os vocabulários que possuam uma semântica bem definida podem proporcionar que o publicador e o consumidor tenham a mesma compreensão do significado contido nos dados (LÓSCIO; BURLE; CALEGARI, 2017).

MP 15 – Reutilizar vocabulários, dando preferência aos padronizados

A utilização de vocabulários amplamente utilizados e padronizados estimula o intercâmbio de dados, diminui as redundâncias, incentiva a reutilização de dados e facilita o consenso entre o publicador e consumidor. Dessa forma, recomenda-se o uso de termos oriundos de vocabulários compartilhados, preferencialmente os padronizados, para codificar dados e metadados. Os benefícios desta prática são o reuso, a processabilidade, a compreensão, a confiança e a interoperabilidade (LÓSCIO; BURLE; CALEGARI, 2017).

MP 16 – Escolher o nível de formalização adequado

A recomendação para esta melhor prática é escolher um nível de semântica formal e adequado que se ajuste tanto aos dados quanto às aplicações mais prováveis de serem utilizadas. Rautenberg, Dall'Agnol e Michelon (2018, p. 47) explicam que:

> A semântica formal ajuda a estabelecer especificações que transmitem um significado preciso e detalhado. É possível fazer uso de vocabulários complexos, os quais podem servir como base para realização de inferências. Por outro lado, vocabulários complexos requerem mais esforços para serem produzidos e entendidos. Dados altamente formalizados são mais difíceis de explorar por motores de inferência. Assim, os produtores de dados devem procurar identificar o nível certo de formalização para domínios particulares, audiência e tarefas.

Segundo Lóscio, Burle e Calegari (2017, on-line), "o resultado dessa prática é que as aplicações mais prováveis sejam suportadas sem um grau de complexidade maior que o

necessário". Dessa forma, obtêm-se os benefícios de reuso, compreensão e interoperabilidade dos dados.

4.4.2.9 Acesso aos Dados

Foram definidas dez (de 17 a 26) melhores práticas relacionadas ao acesso a dados, reforçando a garantia que tanto pessoas quanto máquinas se beneficiem do compartilhamento de dados por meio da infraestrutura Web.

MP 17 – Fornecer *download* em massa (*bulk download*)

A recomendação para esta melhor prática é que a infraestrutura da Web seja implantada de modo a permitir o acesso completo (em massa) de um conjunto de dados com apenas um pedido, evitando inconsistência no acesso individual de dados ao longo de muitas recuperações. O uso de protocolos de transferência de arquivos, do inglês *File-Transfer Protocols* (FTP) garante a transferências de arquivos grandes em menor tempo. Nesta recomendação os benefícios são o reuso e o acesso aos dados.

MP 18 – Fornecer subconjuntos para conjuntos volumosos de dados

Importante ressaltar a possiblidade de permitir ao usuário acessar a subconjuntos de dados, caso não queiram o conjunto completo. Além disso, os conjuntos volumosos de dados são difíceis de armazenar e transferir. Segundo Rautenberg, Dall'Agnol e Michelon (2018, p. 48), é de responsabilidade das "APIs fornecer filtros sobre o conjunto de dados disponível, permitindo que o nível de detalhes seja escolhido de acordo com as necessidades do domínio e da demanda de desempenho do aplicativo da Web". Neste caso, os benefícios são o reuso, o acesso, a interligação e a processabilidade de dados.

MP 19 – Usar negociação de conteúdo para fornecer dados em múltiplos formatos

A negociação de conteúdo permitirá que diferentes recursos ou representações múltiplas do mesmo recurso possam ser recuperados de acordo com a solicitação do usuário. Isto porque, segundo Lóscio, Burle e Calegari (2017, on-line), é possível disponibilizar em uma página HTML dados legíveis por pessoas, mesclados a dados legíveis por máquinas, usando RDFa, por exemplo. Os mesmos dados podem estar disponíveis em JSON, XML, RDF, CSV e HTML. Essas representações múltiplas podem ser disponibilizadas por meio de

uma API, no entanto devem ser disponibilizadas a partir do mesmo URL utilizando-se a negociação de conteúdos. Os benefícios desta aplicação são o reuso e o acesso aos dados.

MP 20 – Fornecer acesso em tempo real

Fornecer dados em tempo real é especialmente importante para a publicação aberta de dados de pesquisa com vista ao avanço da ciência, pois não dependem somente de tecnologias, vocabulários e formatos, mas também da iniciativa do produtor de dados. Contudo, Lóscio, Burle e Calegari (2017, on-line) ponderam que o fornecimento de dados em tempo real para uma determinada aplicação precisa sempre ser avaliado individualmente, considerando as taxas de atualização, a latência introduzida pelos passos de pós-processamento de dados, a disponibilidade de infraestrutura o que os consumidores necessitam. Para os autores, além de disponibilizar os dados, os produtores podem fornecer informações adicionais, tais como a descrição de lacunas, erros e anomalias, bem como informações sobre atrasos de publicações. Os benefícios das aplicações serão a capacidade de acessar os dados em tempo real ou quase em tempo real (intervalo de milissegundos), após a criação dos dados e a disponibilidade para o seu reuso.

MP 21 – Fornecer dados atualizados

A orientação é disponibilizar dados na Web, rigorosamente na data de criação ou coleta, eventualmente após estes terem sido processados ou alterados. Esta orientação oferece os benefícios de acesso e reuso dos dados.

MP 22 – Fornecer uma explicação para os dados que não estão disponíveis

Fornecer uma explicação on-line contextualizando as razões pelas quais os dados não estão disponíveis ou se estão disponíveis sob diferentes condições. Neste caso, os benefícios são o reuso e a confiabilidade.

MP 23 – Tornar os dados disponíveis usando uma API

O uso de uma API oferece mais flexibilidade e capacidade de processamento para os consumidores de dados. Uma API ativa o uso de dados em tempo real, realiza filtragens a partir de solicitações e permite trabalhar com eles em um nível atômico. Na interpretação de Silva, Santarem Segundo e Silva (2018), no caso de dados de pesquisa, o tempo dependerá do tipo de pesquisa que está sendo realizada, enquanto que poderá não ocorrer o acesso a dados

de uma publicação ampliada em tempo real, em função dos trâmites institucionais das defesas públicas, avaliações das bancas avaliativas e depósitos em repositórios. Esta recomendação proporciona aos desenvolvedores o acesso aos dados para uso em seus próprios aplicativos, com dados atualizados e sem necessidade de esforço por parte dos consumidores. Os benefícios desta prática são o reuso, a processabilidade, a interoperabilidade e o acesso aos dados (LÓSCIO; BURLE; CALEGARI, 2017).

MP 24 – Usar padrões Web como base para construção de APIs

As APIs que são construídas sob padrões da Web, tais como *REpresentational State Transfer* (REST) aproveitam os pontos fortes da infraestrutura da Internet, por exemplo, usam verbos HTTP como métodos e URLs que se orientam diretamente a recursos individuais, ajudando a evitar um forte acoplamento entre as requisições e as respostas. Os benefícios desta prática são o reuso, a interligação, a interoperabilidade, a descoberta, o acesso e a processabilidade (LÓSCIO; BURLE; CALEGARI, 2017) e (RAUTENBERG; DALL'AGNOL; MICHELON 2018).

MP 25 – Fornecer documentação completa para as APIs

Segundo Lóscio, Burle e Calegari (2017, on-line), a documentação de uma API é essencial para a sua utilidade e "os desenvolvedores serão capazes de obter informações detalhadas sobre cada chamada para a API, incluindo os parâmetros que leva e o que é esperado para retornar, isto é, todo o conjunto de informações relacionadas com a API". Nesse conjunto de informações sobre os dados são incluídos como usá-lo, avisos de mudanças recentes, informações de contato, dentre outros. Lóscio, Burle e Calegari (2017, on-line) ressaltam que toda a documentação deve estar disponível de modo navegável na Web, a qual permitirá também que as máquinas possam acessar e colaborar na criação de softwares clientes da API. Para esta prática os benefícios são o reuso e a confiabilidade dos dados.

MP 26 – Evitar alterações que afetem o funcionamento de sua API

O consórcio W3C recomenda evitar alterações em sua API as quais afetem o código do cliente e, quando houver evolução, informar seus desenvolvedores sobre quaisquer modificações feitas. Sobre isso, Lóscio, Burle e Calegari (2017, on-line) explicam que o código do desenvolvedor deve continuar válido após alterações na API. Nas palavras de Rautenberg, Dall'Agnol e Michelon (2018, p. 52)

> quando os desenvolvedores implementam uma aplicação cliente de uma API, estes projetam características específicas, como o esquema ou o formato de uma resposta. Evitar alterações significativas na API minimiza a ruptura do código da aplicação cliente. Entretanto, quando isso não é possível, as mudanças devem ser comunicadas aos desenvolvedores para aproveitar as novas características ou realizar ações corretivas em suas aplicações clientes.

Os benefícios proporcionados por esta prática são o reuso e a interoperabilidade.

4.4.2.10 Preservação dos Dados

As melhores práticas 27 e 28 oferecem recomendações para preservar aqueles dados que não são mais usados, bem como apresentam as medidas necessárias para indicar que eles foram retirados ou arquivados. Segundo Lóscio, Burle e Calegari (2017, on-line) "simplesmente apagar um recurso da Web é uma prática ruim". Nesse caso, os autores recomendam utilizar um URI que leva a um código de resposta.

MP 27 – Preservar os identificadores

Esta melhor prática recomenda que ao remover dados da Web, que seja feita a preservação do identificador e forneça informações sobre o recurso arquivado. A resposta esperada desta prática é garantir que o URI de um recurso faça referência ao conjunto de dados ou redirecione para uma informação a respeito. Os benefícios são o reuso e a confiabilidade (LÓSCIO; BURLE; CALEGARI, 2017).

MP 28 – Avaliar a cobertura do conjunto de dados

A preservação de um conjunto de dados particular deve envolver a preservação de todo seu contexto na Web de dados. Rautenberg, Dall'Agnol e Michelon (2018, p. 54) explicam que, "no momento do arquivamento, deve ser realizada uma avaliação da ligação de *dump* do conjunto de dados com os recursos preservados e os vocabulários utilizados". Espera-se nesta prática que, no futuro, os usuários sejam capazes de fazer uso de dados arquivados. Os benefícios desta aplicação são o reuso e a confiabilidade.

4.4.2.11 *Feedback*

O *feedback* proporciona benefícios para os provedores e para os consumidores de dados, tanto para a avaliação da qualidade dos dados publicados como para proporcionar interação entre usuários e provedor. O *feedback* ajuda os publicadores a melhorar a

integridade dos dados, assim como incentivá-los a publicar novos dados. O *feedback* permite ao usuário relatar sua experiência de uso, preferências e necessidades (LÓSCIO; BURLE; CALEGARI, 2017).

MP 29 – Coletar *feedback* dos consumidores de dados

Recomenda-se fornecer uma canal de coleta de *feedback* que seja fácil de encontrar e intuitivo. Para Lóscio, Burle e Calegari (2017, on-line) da perspectiva da interface do usuário, existem diversas maneiras para coletar *feedbacks* de consumidores de dados, incluindo o registro no site, formulários de contato, seleção de avaliações de qualidade, pesquisas e caixas de comentário. Esta prática possibilita que os usuários sejam capazes de fornecer *feedbacks* e avaliações sobre conjuntos de dados e distribuições. Os benefícios proporcionados são o reuso, a compreensão e a confiabilidade dos dados.

MP 30 – Compartilhar o *feedback* disponível

Sempre que possível, os *feedbacks* devem ser disponibilizados publicamente para que outros usuários de dados possam examiná-los. O compartilhamento dos *feedbacks* permite que as pessoas avaliem as experiências de outros, possibilita experiências comunitárias, incentiva um ambiente colaborativo e possibilita que suas experiências comunitárias, preocupações ou perguntas sejam sempre atendidas. Nesta prática, os benefícios são o reuso e a confiabilidade dos dados (LÓSCIO; BURLE; CALEGARI, 2017).

4.4.2.12 Enriquecimento dos Dados

O enriquecimento de dados permite agregar valores omissos aos dados existentes e melhorar a qualidade em sua recuperação. De acordo com Lóscio, Burle e Calegari (2017), o enriquecimento de dados ocorre a partir de um conjunto de processos que pode ser utilizado para aperfeiçoar, aprimorar ou melhorar dados brutos ou dados previamente processados. Os autores ressaltam sobre os cuidados a serem tomados para evitar que o enriquecimento distorça resultados ou conclusões estatísticas, assim como evitar combinação de dados que comprometa a privacidade de pessoas. Esta seção fornece algumas recomendações a serem seguidas por publicadores de dados com a finalidade de enriquecê-los.

MP 31 – Enriquecer dados por meio da geração de novos dados

Publicar conjuntos de dados mais completos pode aumentar a confiança, quando feita de maneira adequada e ética. Sendo assim, novas atribuições e mensurações podem ser atribuídas para dados omissos, a partir de dados brutos pré-existentes. Da mesma forma, dados com valores nulos poderão ser corrigidos a partir do preenchimento de tais valores. Conjuntos de dados também podem ser enriquecidos por meio de coletas adicionais da mesma forma como os dados originais foram coletados, ou combinando dados originais com outros conjuntos. O resultado pretendido com esta prática é a melhoria de conjuntos de dados com dados faltantes, por meio do preenchimento de tais valores. A estrutura poderá ser conferida aos dados e sua utilidade poderá ser melhorada, se forem adicionadas medidas ou atributos relevantes. Porém, tal adição só deverá ser feita se não alterar os resultados analíticos e o significado deles. Os benefícios desta prática são o reuso, a compreensão, a confiabilidade e a processabilidade (LÓSCIO; BURLE; CALEGARI, 2017) e (LÓSCIO; BURLE; OLIVEIRA; CALEGARI, 2018).

MP 32 – Fornecer apresentações complementares

Esta prática visa fornecer informações complementares para que os consumidores humanos possam compreender o conteúdo do conjunto de dados de forma imediata sem ter que criar suas próprias ferramentas. Uma maneira seria a publicação de sumário analítico em uma página HTML; gráficos e tabelas podem colaborar com os usuários ao explorar o resumo para compreender os dados. Neste caso, os benefícios são o reuso, a compreensão, o acesso e a confiabilidade.

4.4.2.13 Republicação

Sobre a republicação Lóscio, Burle e Calegari (2017) explicam que reutilizar dados é outra forma de publicar dados ou simplesmente republicá-los. A republicação de dados inclui a combinação de dados existentes com outros conjuntos, criar aplicativos Web, reempacotamento de dados em uma nova forma, como por exemplo uma tradução linguística dos dados. Para os autores, a republicação de dados apresenta responsabilidades que são exclusivas para essa forma de publicação na Web. As melhores práticas 33, 34 e 35 fornecem recomendações que devem ser observadas para a republicação.

MP 33 – Fornecer *feedback* para o provedor original

Realizar comentários sobre a utilidade dos dados é uma maneira de informar o publicador original sobre a qualidade de seus dados e colaborar para justificar a aplicação de recursos no seu lançamento. Fornecer *feedback* recompensa os publicadores por seus esforços e os ajuda a melhorar seu conjunto de dados para futuros usuários. Além disso, esta prática proporciona clareza das medidas que podem ser adotadas para melhorar os dados. Os benefícios pretendidos com estas ações são o reuso, a interoperabilidade e a confiabilidade dos dados.

MP 34 – Seguir os termos de licença

Esta recomendação indica encontrar e seguir os requisitos da licença informada pelo publicador original do conjunto de dados. Lóscio, Burle e Calegari (2017) destacam que a licença fornece uma estrutura jurídica para utilizar o trabalho de outra pessoa. Dessa forma, o resultado pretendido desta prática é que os publicadores de dados possam confiar que os seus trabalhos estejam sendo utilizados de acordo com as condições da licença, o que provavelmente os estimularão a continuar publicando. Os benefícios adquiridos com a aplicação desta prática são o reuso e a confiabilidade.

MP 35 – Citar a publicação original do conjunto de dados

Ao republicar dados é necessário citar a fonte. A citação da publicação original dos dados informa ao usuário final quem realmente são os criadores de tal conjunto de dados, gera confiabilidade e ajuda o publicador dando crédito ao seu trabalho. A citação também mantém a procedência e ainda ajuda outros a trabalharem. O resultado esperado nesta prática é que os consumidores finais sejam capazes de avaliar a confiabilidade dos dados e que os esforços dos provedores originais sejam reconhecidos. Nesse sentido, os benefícios proporcionados por esta prática são o reuso, a descoberta e a confiabilidade.

Estas melhores práticas se aplicam para a publicação de dados abertos e não abertos. A publicação a partir da abordagem *Linked Open Data* é realizada sob uma licença aberta. A publicação de dados abertos e conectados pode validar a pesquisa e colaborar para a geração do conhecimento científico. Para tanto, os dados precisam ser publicados seguindo orientações e tecnologias amplamente utilizadas. Sendo assim, as tecnologias semânticas são relevantes e, portanto, serão discutidas no capítulo 5 desta tese.

Para propor um conjunto de diretrizes semânticas para a estruturação e a publicação de dados de pesquisa científica anotados em cadernos de pesquisa, faz-se necessário conhecer as características e as especificidades desses tipos de dados. Dessa forma, o capítulo 5 abordará o estudo dos dados de pesquisa de cadernos de laboratórios.

5 CADERNO ABERTO DE PESQUISA

Este capítulo apresenta o Caderno Aberto de Pesquisa ou *Open Notebook Science*, versão original em inglês, conhecido pela sigla ONS, para se referir à técnica de disponibilizar, em formato aberto, os dados de pesquisa científica registrados em cadernos de pesquisa ou cadernos de laboratório, para que outros pesquisadores possam acessar, usar e reutilizar sem restrição, estando sujeito, no máximo, a atender exigências que visem preservar a proveniência e a abertura dos dados. Para contemplar o arcabouço teórico sobre os cadernos abertos de pesquisa, este capítulo foi subdivido nas seções: 5.1 aspectos conceituais e históricos; 5.2 cadernos abertos e a comunidade científica; e 5.3 tecnologias do caderno aberto de pesquisa.

Na seção 5.1, o caderno aberto de pesquisa será descrito em seus aspectos conceituais e históricos resgatando o propósito de sua origem e destacando a trajetória de seu entusiasta Jean-Claude Bradley, bem como as ações desenvolvidas em parceria com os seus colaboradores para fortalecer o movimento de disponibilizar dados gratuitos e em tempo real.

Na seção 5.2 são apresentadas as vantagens que a abertura e o compartilhamento de dados científicos anotados em cadernos de pesquisa podem proporcionar à comunidade científica e à sociedade como um todo.

Na seção 5.3, as tecnologias aplicadas aos cadernos de pesquisa são apresentadas, partindo da perspectiva das três tecnologias: material, social e literária, definidas por Shapin e Shaffer (1985) e analisadas por Clinio (2016) como uma nova tecnologia literária para um novo formato de produzir e comunicar a ciência. São apresentadas ferramentas tecnológicas adotadas em cadernos abertos de pesquisa, com destaque para os projetos *UsefulChem* por ser a primeira iniciativa de caderno, *LabScribbles* e o *Openlabnotebooks* por possuírem maior destaque na literatura da área. Ainda nesta seção serão analisada a estrutura dos dados de pesquisa científica registrados nos cadernos de pesquisa estudados nesta tese.

Em defesa do movimento de produzir e publicar dados abertos, Jean-Claude Bradley, o principal autor de *Open Notebook Science* possui a maioria de suas publicações em *blogs*. Dessa forma, as principais fontes de informação para a realização deste capítulo, em âmbito internacional, foram vídeos, aulas, palestras, entrevistas, capítulos de livros e artigos científicos de autoria de Bradley e seus colaboradores, disponibilizados, na maioria das vezes, em *blogs*. No Brasil, a principal fonte de informação adotada foi a tese de Anne Clinio e capítulos de livros e artigos científicos de autoria de Anne Clinio e Sarita Albagli.

5.1 ASPECTOS CONCEITUAIS E HISTÓRICOS

O termo Caderno Aberto de Pesquisa ou *Open Notebook Science*, foi cunhado por Jean-Claude Bradley, em setembro de 2006, para promover debates sobre a colaboração aberta na ciência e desenvolver técnicas de pesquisas mais eficazes. Para Bradley (2010, tradução nossa), o caderno aberto de pesquisa pode ser definido como "uma nova maneira de fazer ciência na qual - o melhor que puder - você torna toda a sua pesquisa aberta ao público em tempo real" "com o objetivo de atrair colaboradores e recursos para promover uma ciência 'mais rápida, uma ciência melhor' (*fast science, better science*)". (CLINIO, 2016, p. 87).

A proposta de abertura de conjuntos de dados de pesquisa registrados em cadernos de laboratório faz parte de um movimento maior da Ciência Aberta denominado *e-Science*, caracterizado pelo uso intensivo de tecnologias e esforços colaborativos, os quais trazem a oportunidade de se pensar os novos contextos e práticas científicas. Para Clinio e Albagli (2017), a proposta trata-se de uma inovação no modo de produzir e comunicar a ciência, desenvolvida por diversos cientistas entre os quais se destaca Jean-Claude Bradley como o principal entusiasta no recrutamento de apoiadores para a definição do conceito e práticas de cadernos abertos em química.

Os cadernos abertos de pesquisa foram considerados por Bradley (2010) uma prática de tornar os detalhes de todos os experimentos feitos em seu laboratório disponíveis gratuitamente na Web, o que não se limitava apenas a uma descrição, mas também incluem todos os dados gerados a partir de experimentos, até mesmo os experimentos mal sucedidos. Bradley (2010) ressalta que o objetivo dos cadernos abertos é tornar os dados brutos, em vez de pesquisas publicadas, as quais estariam disponíveis gratuitamente dentro de horas de produção, e não após meses ou anos envolvidos na revisão de pares. Bradley, Lang, Koch e Neylon (2011) complementam que a publicação dos dados de pesquisa, em tempo real, permite que outros pesquisadores contribuam rapidamente, pois se pode supor que, se um experimento não for relatado, ainda não foi feito. Os colaboradores em potencial podem, então, realizar de maneira confiável os experimentos sem se preocuparem com a necessidade de duplicar o trabalho. Se eles decidirem replicar uma experiência, podem fazê-lo com o conhecimento prévio do que aconteceu em todas as tentativas anteriores.

Na interpretação de Clinio (2016, p. 87),

> o caderno aberto é uma prática que, em seu nível ideal, disponibiliza a íntegra dos registros individuais de um pesquisador ou de um conjunto de cientistas de um laboratório, on-line e em tempo real, através de licenças livres que permitem o acesso, reutilização e redistribuição do conteúdo por qualquer pessoa. Esta prática não inclui apenas dados, informações e

> resultados favoráveis de uma pesquisa científica, mas compartilha também status parciais, debilidades e desafios quando eles ainda não foram resolvidos pelos pesquisadores.

De acordo com o Open Notebook Network (ONSNetwork)[9] (2019?), o caderno aberto de pesquisa é

> simplesmente a prática de disponibilizar todo o projeto de pesquisa on-line à medida que é registrado. Esse local on-line é conhecido como um bloco aberto de anotações e é semelhante ao caderno de papel que a maioria dos cientistas mantém em seu laboratório. É o centro de armazenamento de planos de projetos, protocolos e configurações experimentais, dados brutos e até mesmo interpretações não filtradas. Um caderno pode ser qualquer tipo de site, desde que atenda às necessidades do cientista e esteja disponível publicamente (algumas plataformas de cadernos úteis). (ONSNetwork, 2019?, on-line, tradução nossa).

Ao pensar a nova prática de disponibilizar dados brutos anotados em cadernos de laboratório, Bradley (2006) buscou definir um termo que apresentasse o conceito de sua proposta, diferente dos termos relacionados à *Open Source Software* e *Open Source Science* [Ciência de Código Aberto], para que não fosse confundido com outras práticas. Bradley (2006) relatou que a definição de ciência de código aberto foi considerada, a princípio, consistente com a proposta de seu projeto

> Na ciência de código aberto, o código é disponibilizado para qualquer pessoa para modificar e reutilizar. O que temos tentado fazer com o *UsefulChem* [seu projeto] é fornecer a entidade análoga para a pesquisa química, que é um dado experimental bruto, juntamente com a interpretação do pesquisador, em um formato que qualquer um pode facilmente re-analisar, reinterpretar e reutilizar. Um bom exemplo de reutilização é o uso de alguns resultados e observações de um experimento fracassado de uma maneira que nunca foi planejada pelo pesquisador original. Isso simplesmente não acontece regularmente na ciência porque os experimentos fracassados quase nunca são incluídos nas publicações. (BRADLEY, 2006, on-line, não paginado, tradução nossa). (Publicado no blog em 26 de setembro de 2006).

Porém, o termo ciência de código aberto é usado em abordagens mais restritas como, por exemplo, em discussões sobre *pré-prints* de artigos em periódicos, ou seja, artigos com os resultados de pesquisa que não foram publicados ainda em um período científico com revisão por pares. A proposta de Bradley se distancia ao demandar a publicação de dados brutos. Assim, nomeou a nova prática de *Open Notebook Science*, com a seguinte definição

> Para esclarecer a confusão, usarei o termo *Open Notebook Science*, que ainda não sofreu mutação memética. Com ela, eu me refiro à existência de

[9] A rede Open Notebook Science Network (ONSNetwork) e seus cadernos são construídos em *Word Press*, personalizados para uso acadêmico pela Open Science Federation, liderada por Brian Glanz e com orientação de cientistas de cadernos abertos Anthony Salvagno e Jean-Claude Bradley.

> uma URL linkada a um caderno de laboratório que está disponível abertamente e indexado em mecanismos de pesquisa comuns. Ele não precisa necessariamente parecer um caderno de papel, mas é essencial que todas as informações estejam disponíveis para os pesquisadores tirarem suas conclusões e estejam igualmente disponíveis para o resto do mundo. Basicamente nenhuma informação privilegiada. (BRADLEY, 2006, on-line, não paginado, tradução nossa) (Publicado no blog em 26 de setembro de 2006).

O conceito de caderno aberto foi inserido nas atividades de Bradley e de outros adeptos ao movimento com a justificativa de que o caderno aberto pode acelerar e melhorar a qualidade da ciência. Dessa forma, Bradley compartilhava não apenas os dados e informações geradas em suas pesquisas, mas também compartilhava informações que influenciava as condições de realização do seu trabalho. Bradley interagia com os pesquisadores de forma pública, por meio de *blogs* e *wikis*, com sugestões de leituras de outros colaboradores, acompanhava propostas de colegas e sugeria melhorias nas pesquisas de colegas. Além disso, Bradley dispõe de todo o seu material de aula disponível em *blogs* (CLINIO, 2016).

Em uma entrevista a Richard Pointer, revista *Information Today*, no ano de 2010, Bradley expôs que, como a maioria dos cientistas mantinha sua pesquisa em sigilo até a publicação em seu formato final, como artigos e capítulos de livros, e solicitou patentes de seu trabalho em nanotecnologia e terapia gênica, porém identificou que a cultura do sigilo é em parte responsável por não proporcionar o impacto esperado nas pesquisas científicas. Esse impacto pode estar relacionado ao tempo entre a realização do experimento e a publicação final, a falta de detalhamento dos dados e pela ausência de divulgação dos experimentos mal sucedidos. Portanto, aderiu à abordagem de abertura de dados brutos, sejam em gráficos, textos, números, fórmulas e outros, que permite a conexão e colaboração entre cientistas.

Em 2005, Bradley lançou uma iniciativa baseada na Web, chamada de *UsefulChem*, como uma tentativa de ciência de código aberto na química. O projeto *Useful Chemistry* ou *UsefulChem* visou trazer problemas importantes e globais à atenção da comunidade de química mais ampla, na esperança de encontrar soluções baseadas na química. Bradley (2010) relatou em seu primeiro *post* que pretendia trabalhar com algo que pudesse fazer, sendo um químico, de necessidade urgente para a sociedade. Então, começou a pesquisar por frases em artigos de química que respondessem a uma necessidade urgente. Assim, obteve de suas buscas a ideia de trabalhar com a malária, tema potencialmente muito útil, pois a malária mata milhões de pessoas todos os anos, pessoas estas que vivem no mundo em desenvolvimento, onde as grandes empresas farmacêuticas não estão dispostas a dedicar-se muito tempo ao desenvolvimento de novos tratamentos. Nesse sentido, a proposta de Bradley (2010) recai

para a abertura de dados científicos que acelerassem descobertas e combatessem a malária. Dessa forma, Bradley (2010) publica os detalhes de todos os experimentos feitos em seu laboratório, deixando-os disponíveis abertamente. Ele não limita isso apenas a uma descrição, mas também inclui todos os dados desses experimentos, até mesmo os menos sucedidos. Segundo o autor, é essencial que todas as informações estejam disponíveis para os pesquisadores tirarem suas conclusões, igualmente disponível para o restante do mundo. Para Harding (2019), essa prática significa que o caderno deve ser uma representação completa e honesta das descobertas dos cientistas.

Para elucidar os objetivos de uso do caderno aberto de pesquisa, Bradley (2009) apresenta as propriedades: reprodutibilidade e exclusão. A primeira propriedade apresenta o principal objetivo, que é o de fazer experiências reprodutíveis para o pesquisador que a registrou, pois o pesquisador não poderá melhorar um processo se não mantiver um registro detalhado do que exatamente ocorreu em um determinado julgamento. Para Bradley (2009), um propósito secundário é provar o que um pesquisador sabia e o que fazia em um determinado momento específico. Assim, o caderno deve conter todas as informações necessárias para reproduzir os resultados obtidos. Isso significa que, quando é publicado perto do tempo real, outros pesquisadores que não conhecem o projeto, podem ler os detalhes de uma experiência feita hoje e podem colaborar para o avanço desse projeto amanhã.

Na propriedade exclusão, o caderno aberto permite que um pesquisador conheça os detalhes do projeto e experimentos realizados por outro pesquisador, mesmo que mal sucedidos. Ou seja, se uma determinada pesquisa não for localizada, significa que outro pesquisador possa desenvolvê-la. O objetivo do caderno aberto é obter toda a verdade sobre o que foi feito e o que ainda não foi em determinada pesquisa, pois segundo Bradley (2009, não numerado, tradução nossa) “se eu não encontrar o que estou procurando no seu caderno – e você declarou que ele é um caderno aberto – então eu posso seguramente assumir que você não fez isso e eu vou sentir confiante em investir meus recursos para fazer o próximo experimento”. Além disso, a partir do acesso aos detalhes dos experimentos registrados no caderno, o pesquisador poderá contribuir para o desenvolvimento da pesquisa, desde que conheça os detalhes de onde, como e quando estão sendo realizados os experimentos e os resultados parciais obtidos, inclusive os “mal sucedidos”, como mencionado, para evitar a repetição das mesmas tentativas, ou repetir uma das falhas, por poder considerar um parâmetro ou erro na análise.

5.1.1 Partial Open Notebook Science (PONS)

Os benefícios da prática de disponibilizar, em formato aberto, dados de pesquisa registrados em cadernos de laboratório, o mais próximo do tempo real, sem restrições ou omissões, somente serão alcançados em sua totalidade quando estes critérios forem atendidos. Nesse contexto, Bradley (2009) menciona que quando discute sobre caderno aberto de pesquisa está se referindo ao

> "fornecimento do acesso a todo o caderno de laboratório e associados a dados brutos em quase tempo real. Como expliquei nesse *post*, somente quando esses dois critérios são atendidos você obtém o beneficio total que qualquer um [usuário] (humano ou não) possa ver o que você fez e não fez e pode realisticamente contribuir com o próximo passo do projeto sem ter que entrar em contato com o pesquisador para mais informações". (BRADLEY, 2009, não paginado, tradução nossa, grifo nosso).

Porém, a abertura total dos dados não é de interesse de todos os pesquisadores, especialmente, quando a proteção à propriedade intelectual está envolvida. Para esses casos, Bacon (2008) sugere formas mais restritivas de abertura de cadernos pesquisa (ONS) e as nomeias de Pseudo-ONS, pseudo cadernos abertos de pesquisa. Enquanto que Bradley (2009) sugere que as chamem de *Partial-ONS* (PONS), abertura parcial de cadernos de pesquisa. Assim, inspirados pelo sistema *Creative Commons*, Bradley (2009) com a colaboração de Hope Leman e Andy Lang criaram uma declaração do ONS para vincular os logos às descrições, conforme o nível de abertura dos dados. As opções de logo são ACI (*All Content – Immediate* / Conteúdo total – imediato), ACD (*All Content – Delayed* / Conteúdo total – Com atraso), SCI (*Selected Content – Immediate* / Conteúdo selecionado - Imediatamente). Apresenta-se a seguir as descrições e os logos dos níveis de abertura do caderno de pesquisa.

Quadro 11 - Logos e graus de abertura do caderno aberto de pesquisa

Nível de abertura do caderno de pesquisa	Descrição
ACI All Content – Immediate	**Todo conteúdo, imediatamente** - a íntegra do caderno de pesquisa e os dados brutos associados à pesquisa estão disponíveis para o público em intervalo de tempo próximo ao real. Se algo não está publicado, outras pessoas podem assumir que o cientista não fez.
ACD All Content – Delayed	**Todo conteúdo, com atraso** - A íntegra do caderno de pesquisa e dos dados brutos associados estão disponíveis, porém com atraso significativo na sua publicação – provavelmente por conta de patenteamento e publicação.

SCI Selected Content – Immediate	**Conteúdo selecionado, imediatamente** - Apenas uma parte do caderno de pesquisa e dos dados brutos associados à pesquisa estão disponíveis e são publicados em um intervalo de tempo próximo ao real. Outros não devem supor que algo não publicado não foi realizado pelo cientista.
SCI Selected Content – Delayed	**Conteúdo selecionado, com atraso** - Apenas uma parte do caderno de laboratório e dos dados brutos associados à pesquisa estão disponíveis e com atraso. Outros não devem supor que algo não publicado ainda não foi realizado.

Fonte: Adaptado de Bradley (2009) e Clinio (2016, p. 94).

Segundo o Open Notebook Science Network (2019?), o ideal seria que os cientistas abrissem todos os aspectos de sua pesquisa em tempo real, mas existem fatores que dificultam esse movimento como, por exemplo, receios de lidar com acesso aberto completo, conflitos com propriedade intelectual e sobrecarga de publicações de dados on-line. Nesse sentido, a escala de logos e níveis apresenta diferentes possibilidades de aberturas de cadernos de pesquisa, mesmo que sejam projetos antigos. Bacon (2008) defende que publicar etapas do projeto em atraso, por exemplo, um mês de atraso do seu caderno de pesquisa atual, viola o propósito do caderno aberto, mas abre a sua pesquisa para outros interessados acompanharem.

A PONS e suas definições de abertura de dados não são vistas por todos os pesquisados e projetos, Todd (2011), coordenador do projeto descoberta de drogas de código aberto para a malária, estabelece seis princípios básicos para quem tiver interesse em participar do projeto, são eles:

1 – Primeira Lei: todos os dados são abertos e todas as ideias são compartilhadas

2 – Segunda Lei: qualquer um pode participar em qualquer nível do projeto

3 – Terceira Lei: não haverá patentes

4 – Quarta Lei: Sugestões são a melhor forma de crítica

5 – Quinta Lei: Discussão pública é muito mais valiosa do que e-mail privado

6 – Sexta Lei: O projeto é maior do que qualquer outro laboratório e não pertence a ele. O objetivo é encontrar uma boa droga para a malária, por qualquer meio, o mais rápido possível.

Segundo Todd (2011, on-line) qualquer pesquisador que tenha interesse em colaborar com o projeto poderá fazê-lo em qualquer nível. “Com uma ressalva – que ao participar você entenda que quer que você faça precisa retribuir de volta ao projeto na integra” –. A proposta era a construção colaborativa do projeto por meio do compartilhamento de dados abertos.

Completa Todd (2011, on-line), "[...] por favor, nem pense em participar se você não estiver completamente ciente sobre a necessidade e os benefícios espetaculares de compartilhar todos os dados e ideias em público" (TODD, 2011, on-line, não paginado, tradução nossa).

Para Harding (2019), apesar dos benefícios de documentar abertamente projetos de pesquisa em tempo real, os cientistas demoraram a aderir a prática e aqueles que o fizeram, muitos abandonaram rapidamente ou não atualizam seu caderno regularmente ou o compartilham com restrições.

Até o momento da realização desta pesquisa não foi possível identificar legislações que regulamentam a prática de cadernos abertos. Contudo, é possível observar que "pesquisas sobre doenças raras é especialmente acessível ao modelo de cadernos abertos porque pode aumentar o impacto cientifico e servir como um mecanismo para engajar grupos de pacientes no processo científico" (HARDING, 2019, on-line).

5.1.2 Jean-Claude Bradley: entusiasta do conceito *Open Notebook Science*

O contexto histórico e conceitual dos Cadernos Abertos de Pesquisa está associado com a trajetória de Jean-Claude Bradley - químico, professor, pesquisador e coordenador de *E-learning* na Universidade de Drexel, na Filadélfia, Estados Unidos – que liderou as discussões acerca da Ciência Aberta em Química, foi o autor da proposta de *Open Notebook Science*, desenvolveu estudos químicos de destaque à sociedade, dentre eles os compostos antimaláricos. Ele usava os serviços da internet como *blogs*, *wikis* e outras páginas da Web para divulgar os seus experimentos químicos e torná-los abertos em tempo real.

Para Clinio (2016), Bradley foi o disseminador do conceito de caderno aberto de ciência por meio dos inúmeros documentos produzidos e compartilhados de sua própria prática cotidiana. A prática de compartilhamento de Bradley consistiu na divulgação de seus experimentos em andamento, materiais de aulas, artigos e capítulos de livros, entrevistas e palestras diversas em seus vários *blogs*, dentre eles *UsefulChem*, *Drexel CoAS E-Learning* e no seu caderno aberto em formato *wiki Open Notebook Science*. Os compartilhamentos de Bradley incluíam os dados brutos e os detalhes de como processava os dados, a interpretação dos resultados e os seus planos futuros. A partir desses compartilhamentos outros pesquisadores e estudantes realizavam comentários, sugestões e recomendações aos experimentos proporcionando melhorias ao estudo.

O químico Jean-Claude Bradley foi o principal entusiasta nos debates sobre a abertura da ciência e definição de conceitos e práticas do *Open Notebook Science*. Segundo Clinio (2016), Bradley foi um recrutador de cientistas como Andrew Lang, Bill Hooker, Cameron

Neylon, Rajarshi Guha, Steve Koch, Anthony Williams, Mathew Todd, Anthony Salvagno, Philip Rosenthal, Daniel Zaharevitz, Egon Willighagen, David Bradley, Peter Murray-Rust, entre outros colaboradores para debater sobre os conceitos e práticas de *Open Notebook*. Nessa dinâmica, Bradley elucidava que a ciência se beneficiaria mais com a abertura de dados e encorajava os cientistas a disponibilizarem publicamente todo o registro primário de um projeto de pesquisa on-line, principalmente por meio de ferramentas da Web 2.0, e assim definir a direção para o que se tornará cada vez mais o caminho comum para liberar dados e progressão da ciência para o mundo.

Segundo Bradley (2006), a iniciativa de maior destaque foi a criação do *blog Useful Chemistry* ou *UsefulChem* para identificar problemas importantes e globais à atenção da comunidade química em geral, na esperança de encontrar soluções baseadas nessa área. Para Drexel (2014), o projeto evoluiu para a criação de compostos antimaláricos para ajudar na síntese de medicamentos para combater a malária. Novos projetos de incentivo e divulgação da proposta de publicação aberta de dados experimentais foram desenvolvidos por Bradley, entre eles o *Open Notebook Science Challenge* (*ONSchallenge*), na tradução livre Desafio de Caderno Aberto de Ciência. O *ONSchallenge* foi um projeto de pesquisa *crowdsourcing* que envolveu pesquisadores e alunos do ensino médio dos Estados e Unidos e Reino Unido a coletarem medidas de solubilidade não aquosa de compostos orgânicos e as publicarem no blog *UsefulChem*. Como forma de incentivo aos participantes, o projeto *ONSchallenge* promovia premiações aos autores dos melhores trabalhos por meio de dinheiro e publicação em renomados periódicos da área.

O químico participou da criação da plataforma *ChemSpider*, um banco de dados de estrutura química livre utilizado para armazenar informação e permitir consultas de dados em química. Essa plataforma é de propriedade da Royal Society of Chemistry, uma sociedade científica que visa apoiar a investigação científica na área da química. Segundo Bradley, Lang, Koch e Neylon (2011), a plataforma teria a capacidade de fornecer propriedades experimentais do *UsefulChem* como resultado das buscas. Tempos depois o projeto *UsefulChem* adquiriu um subdomínio no *ChemSpider* para armazenar seus experimentos e realização de *upload* de dados abertos.

Na concepção de Bradley, a abordagem de ciência gratuita e acessível combina com a abertura do ensino, em que dados reais do laboratório podem ser usados em tarefas para praticar conceitos aprendidos em sala de aula. Nesse contexto, o químico adota o caderno aberto no ensino de química orgânica e defende que as tendências em educação aberta, ciência aberta e automação são fatores de mudança no cenário educacional do futuro. Seguindo esse

raciocínio, Bradley adotou a ferramenta *Second Life*, um ambiente virtual e tridimensional que permite a visualização dos experimentos em 3D, no ensino de química orgânica e recomenda o uso de outras ferramentas da Web 2.0, como *blogs*, *wikis*, listas de discussão, nas atividades de ensino (DREXEL..., 2014).

O entusiasta nos debates sobre a abertura da ciência colaborou com a divulgação e conscientização do projeto de abertura da pesquisa cientifica apresentando o *Open Notebook Science* para educadores, bibliotecários e profissionais de outras áreas.

Em reconhecimento ao seu trabalho, Bradley recebeu, em 2007, o prêmio *Blue Obelisk*, referente a um grupo informal de químicos, criado em 2005 por Peter Murray-Rust, com a intenção de promover projetos em dados abertos, códigos abertos e padrões abertos. Bradley ficou reconhecido pela comunidade da Ciência Aberta como um dos mais influentes cientistas abertos de nosso tempo. Em 2014, ele faleceu deixando o seu legado como tema do simpósio "*Defining the Future for Open Notebook Science* – A Memorial *Symposium Celebrating the Work of* Jean-Claude Bradley", realizado em 14 de julho de 2014, na Universidade de Cambridge (CLINIO, 2016).

As iniciativas de Bradley, no percurso de 2006 a 2014, tiveram um efeito profundo sobre como os cientistas passaram a se comunicar uns com os outros, propiciando o aumento da colaboração entre pesquisadores em prol da rápida divulgação de resultados. Sendo assim, na seção 5.2, cadernos abertos e a comunidade científica, serão apresentados os benefícios da prática de publicação de dados abertos para a comunidade científica.

5.2 CADERNOS ABERTOS E A COMUNIDADE CIENTÍFICA

O caderno de laboratório, como uma das diversas tipologias de conjunto de dados de pesquisa, vem recebendo reconhecimento pela comunidade cientifica como parte essencial das boas práticas de pesquisa. A prática de abertura e compartilhamento do caderno de pesquisa – em que os pesquisadores compartilham suas pesquisas, incluindo protocolos detalhados, resultados positivos e negativos, on-line e em tempo real – proporciona o aumento da eficiência da comunicação cientifica global.

Para Tenopir *et al*. (2011), os dados de pesquisa são a infraestrutura da ciência, pois formam a base para boas decisões científicas, gerenciamento e uso de recursos e tomada de decisão informada. A Ciência, no contexto da *e-Science*, está se tornando cada vez mais intensa em dados e ações colaborativas. Nesse contexto, os dados digitais não são apenas os resultados da pesquisa, mas fornecem insumos para novas hipóteses, possibilitando novos *insights* científicos e impulsionando a inovação.

Dada a importância dos dados gerados em pesquisas científicas para o desenvolvimento da ciência, faz-se necessário o compartilhamento para o efetivo acesso, uso e reuso desses dados. Segundo Tenopir *et al.* (2011), dentre as vantagens do compartilhamento de dados de pesquisa científica incluem:

- a re-análise de dados que ajuda a verificar os resultados, que é uma parte fundamental do processo científico;
- diferentes interpretações ou abordagens dos dados existentes contribuem para o progresso científico - especialmente em um ambiente interdisciplinar;
- a preservação bem gerenciada e de longo prazo ajuda a manter a integridade dos dados;
- quando os dados estão disponíveis, a (re) coleta de dados é minimizada; assim, o uso de recursos é otimizado;
- a disponibilidade de dados oferece salvaguardas contra má conduta relacionada à fabricação e falsificação de dados;
- estudos de replicação servem como ferramentas de treinamento para novas gerações de pesquisadores.

Os autores Schapira e Harding (2019) também acreditam que a abertura e o compartilhamento de dados de pesquisa científica registrados em cadernos de pesquisa são uma maneira eficiente e rápida de disseminar dados antes de serem publicados em periódicos revisados por pares e apresentam vantagens em relação ao tradicional (*release after publication*).

> Primeiro ao tornar os dados acessíveis em semanas, em vez de mantê-los ocultos por anos, significa que outros poderão aproveitar a pesquisa e evitar gastar tempo e recursos experimentais redundantes. Em segundo, os cadernos de laboratório aberto devem incluir protocolos detalhados que possam ser reproduzidos, o que frequentemente não é o caso em publicações revisadas por pares. Terceiro, os dados mal sucedidos, que quase nunca são divulgados no atual sistema de publicação, mas são fornecidos em cadernos de laboratório abertos podem, portanto, ajudar a economizar tempo, recursos e conhecimento (SCHAPIRA; HARDING, 2019, p. 3).

Segundo a Rede de Cadernos Abertos de Pesquisa (2019?), nome original em inglês Open Notebook Science Network (ONSNetwork), o caderno aberto pode melhorar a capacidade acadêmica por meio de dois mecanismos principais: (1) redução de erros experimentais, atalhos e falsificação de dados; (2) aumento da eficiência da pesquisa pessoal.

O primeiro mecanismo tem como foco a transparência da pesquisa. Ao manter um registro aberto, o processo científico se torna mais transparente na medida em que o

pesquisador disponibiliza além de resultados prévios, os experimentos, a aquisição e análises dos dados. Assim, os erros experimentais podem ser observados antes da conclusão do projeto. Isto porque, os experimentos mal sucedidos quando identificados pelo pesquisador são descritos nos cadernos abertos para que outros pesquisadores não repitam tais erros, porém quando os erros não são identificados pelo autor, podem ser notados mais cedo no processo de pesquisa, não apenas pelo pesquisador, mas por outros membros do laboratório e outros pesquisadores independentes do laboratório. A ONSNetwork (2019?) reforça que mantendo a transparência dos dados, os cientistas não precisariam repetir erros experimentais e não haveria a necessidade de pesquisar por anotações manuscritas antigas ou projetos anteriores. Além disso, não haverá questionamentos quanto a integridade da pesquisa porque todo o registro estará disponível para qualquer pessoa.

O segundo mecanismo está relacionado com a velocidade na publicação da informação. Considerando que a prática de caderno aberto prevê a publicação em tempo real, ou mais próximo possível do tempo real, dos aspectos de uma pesquisa, os cientistas terão o tempo minimizado para iniciar um experimento ou até mesmo repetir um experimento de outro laboratório. No caso de repetição de experimentos, o tempo para garantir a viabilidade seria reduzido, considerando que já não seria um experimento sujeito a erros de interpretação.

Seguindo a mesma linha de raciocínio, Tenopir *et al*. (2011) destacam os benefícios da abertura dos dados de caderno de pesquisa do ponto de vista do pesquisador, das instituições de ensino superior e das agências de fomento. Para os pesquisadores, o acesso aberto a dados de pesquisa, proporciona maior visibilidade, pois podem ser citados por membros mais proeminentes de seu campo e o pesquisador pode receber amplo destaque na comunicação científica. Enquanto que as instituições de ensino superior são beneficiadas a partir de monitoramento e avaliação das atividades científicas ao mesmo tempo em que proporciona prestígio, visibilidade e democratização do acesso. Do ponto de vista das entidades financiadoras de pesquisa, o depósito em acesso aberto das investigações justifica o investimento, cria transparência e evita duplicação de financiamento. Para a pesquisa, portanto, pode avançar para beneficiar a comunidade científica global (TENOPIR *et al*., 2015).

Ainda segundo Tenopir *et al*. (2015), o acesso aberto a dados de pesquisa foi promovido em bem público, uma medida de economia de custos, uma maneira de aumentar a transparência e permitir a verificação de resultados de pesquisas anteriores. Neste sentido, muitas instituições de fomento à pesquisa consideram necessário que os dados resultantes de

projetos financiados sejam gerenciados e compartilhados de forma a garantir maior beneficio para o avanço científico e tecnológico.

Schapira e Harding (2019) destacam que as agências de financiamento estão visualizando o movimento da ciência aberta como uma mudança duradoura e de longo alcance, e assim fortalecendo os princípios de dados abertos. Dessa forma, os autores mencionam algumas das instituições que fomentam pesquisas, em âmbito internacional, e que tem financiado projetos de pesquisa. Dentre elas, encontram-se a Wellcome Trust[10], o Instituto Canadense de Pesquisa em Saúde[11] (*CIHR*, na sigla em inglês), a Fundação Gates[12] e a Iniciativa Chan-Zuckerberg[13] que patrocinaram e apoiaram a realização do simpósio de lançamento do blog openlabnotebooks.org. Ainda são citados o Instituto Nacional de Envelhecimento do NIH[14] que dedicou uma sessão para abrir a ciência na cúpula de pesquisa sobre Alzheimer de 2018; o Congresso de Enroll-HD da CHDI Huntington's Disease Foundation[15], a Sociedade Huntington do Canadá[16] e a Sociedade Huntington da América[17] as quais financiaram estudos de bioquímica HD no Structural Genomics Consortium (SGC), em Toronto.

O princípio da abertura de dados de pesquisa científica encontra-se no aprimoramento da pesquisa para beneficiar a comunidade científica e a sociedade global. A transparência e a velocidade na disponibilidade dos dados podem manter os pesquisadores atualizados e a ciência estaria operando na vanguarda do pensamento, em vez de ficar para trás. Paralelo a isso, temos a morosidade na atualização da literatura em decorrência de uma variedade de processos na revisão de pares.

---

[10] O Wellcome Trust instituição, sediada no Reino Unido, criada para financiar pesquisas para melhorar a saúde humana e animal. Disponível em: https://wellcome.ac.uk.

[11] CIHR é a principal agência federal de financiamento de pesquisas da saúde do Canadá. Disponível em: http://www.cihr-irsc.gc.ca/e/193.html

[12] A Gates Foundation é uma instituição filantrópica fundada por Bill Gates e sua esposa Melina Gates para trabalhar em combate a desigualdade social. Disponível em: https://www.gatesfoundation.org/

[13] O Institute Chan Zuckerberg é uma empresa limitada criada por Mark Zuckerberg e sua esposa Priscilla Chan com a missão de encontrar maneiras de alavancar a tecnologia e soluções voltadas para a comunidade e colaboração para acelerar o progresso em ciência, educação, justiça e oportunidade. Disponível em: https://chanzuckerberg.com/

[14] NIH's National Institute on Aging é a principal agência federal que apoia e conduz a pesquisa da doença de Alzheimer. Disponível em: https://www.nia.nih.gov/

[15]CHDI Foundation é uma organização de pesquisa com financiamento privado unicamente para curas da doença de Huntington. Disponível em: https://chdifoundation.org

[16] Huntington Society of Canada tem como missão os estudos focados na melhoria da qualidade de vida dos afetados pela doença de Huntington. Disponível em: https://www.huntingtonsociety.ca

[17] Huntington's Disease Society of America é uma organização sem fins lucrativos com o objetivo de descobrir a cura para a doença de Huntington. Disponível: https://hdsa.org

5.3 TECNOLOGIAS DO CADERNO ABERTO DE PESQUISA

As tecnologias adotadas pelos cadernos de pesquisa para publicar e compartilhar seus dados experimentais causou grandes mudanças no modo de fazer e comunicar a ciência contemporânea. Segundo Bradley (2009) esclarece que, no passado, os pesquisadores não poderiam tornar as informações públicas em função da ausência de um veículo de publicação conveniente, sendo necessário aguardar a publicação em formato impresso. Atualmente, a tecnologia favorece e existem serviços de hospedagem gratuitos de alta qualidade que permitem o compartilhamento de dados de pesquisa. No âmbito dos cadernos de pesquisa, Bradley (2009) orienta a registrar os cadernos em mídias que são fáceis de compartilhar e indexados automaticamente nos principais motores de busca.

No cenário atual, o uso intensivo de tecnologias na publicação de dados de pesquisa tem afetado a velocidade com que a ciência pode progredir e nos tipos de colaboração são possíveis. Nesse sentido, observa-se na literatura abordagens que favorecem a publicação e o compartilhamento de dados de pesquisa registrados em cadernos de laboratório. Curty (2017) destaca as abordagens dos repositórios de dados, publicações ampliadas e artigos de dados para publicação, de modo a potencializar o valor intrínseco e a autonomia dos dados científicos.

Os repositórios de dados são serviços on-line que podem ser institucionais, temáticos, ligados a comunidades disciplinares ou a projetos de pesquisa. Segundo Curty (2017, p. 5) "os repositórios de dados são responsáveis por reunir, descrever, e promover o acesso e a preservação em longo prazo a conjuntos de dados". Dentre as plataformas de repositórios de dados citados por Curty (2017) incluem: o re3data, o *Global Biodiversity Information Facility* (GBIF)[18] e a *Artic Data Center*[19] como ferramentas que auxiliam e promovem o reuso de dados. Como mencionado no item 3.2.1 (dimensão tecnológica da e-*Science*), o re3data atua como um registro global de repositórios de dados de pesquisa que abrange repositórios de dados de pesquisa de diferentes áreas do conhecimento; oferece repositórios para armazenamento permanente e acesso a conjuntos de dados para pesquisadores, agências de fomento, editores e instituições acadêmicas. A plataforma GBIF fornece ferramentas de busca, de visualização e de exportação de dados; adota padrões de dados que fornecem acesso aberto aos seus conjuntos de dados e atribuição da licença *Creative Commons* para o reuso de cada conjunto. A plataforma *Artic Data Center* reúne e gerencia coleções de dados resultantes

[18] *Global Biodiversity Information Facility* (GBIF). Disponível em: https://www.gbif.org/
[19] *Artic Data Center*. Diponível em: https://arcticdata.io/

de pesquisas relacionadas à região Ártica; adota tecnologias abertas e permite a atribuição de identificadores persistentes para controle de autoridades.

Além das mídias sociais como *blogs* e *wikis*, os repositórios de dados têm sido adotados como plataforma para publicar e compartilhar dados de pesquisa, conforme será apresentado no decorrer dos itens desta seção. Apesar da adoção de repositórios de dados por diferentes cadernos de pesquisa, Curty (2017, p. 17) conclui que essa abordagem "oferece menos atrativos do ponto de vista da lógica do reconhecimento do trabalho científico", sugerindo que "os pesquisadores tendem a preferir os meios e formatos de disseminação cientifica que gerem maior visibilidade e que seja convertido em mais créditos e citações".

A abordagem da publicação ampliada (*enhanced publications*) possibilita a interligação entre manuscritos e dados de pesquisa. Segundo Verhaar (2008, p. 7) "uma publicação pode ser ampliada a partir da adição de um ou mais recursos a um *e-Print*". Esses recursos podem ser constituídos pelos dados (gráficos, imagens, números, som etc.) que sustentam os resultados da pesquisa. Nesse sentido Sales (2014, p. 8) expõe que "o objetivo da publicação ampliada é ligar os resultados de pesquisa aos dados que o geraram, extrapolando o limite do documento físico". Dessa forma, Curty (2017) acredita que a agregação de conjunto de dados possibilita a reutilização para novas pesquisas e reinterpretação dos dados em diferentes contextos, além de proporcionar maior transparência e possibilidade de verificação dos mesmos e maior interatividade dos métodos de revisão por pares, dentre outras vantagens. Ainda assim, a pesquisa de Curty (2017, p. 17) indica que "esse modelo de publicação também não permite a citação de dados de modo independente, dada a inexistência de identificadores e metadados apropriados".

Os artigos de dados (*data papers*) são, na interpretação de Curty (2017, p. 10), "aqueles que buscam descrever uma coleção ou coleções de dados de pesquisa, sem que se estendam à interpretação e inferências dos mesmos". O estudo de Curty (2017, p. 17) aponta que os artigos de dados (*data papers*) são considerados os mais apropriados para publicação de dados de pesquisa científica, pois "descrevem exaustivamente a coleção de dados, acompanhado de descrições do contexto, do percurso metodológico e dos aspectos procedimentais da pesquisa, e das possíveis aplicações dos dados".

A partir das abordagens apresentadas é possível observar que as tecnologias aplicadas ao novo formato de publicação de dados vêm inovando e dinamizando o ecossistema da comunicação científica, por meio da reconfiguração e da sofisticação dos métodos para o compartilhamento e o reuso de resultados de pesquisas.

No entanto, para Clinio e Albagli (2017, p. 7) as inovações proporcionadas aos cadernos de laboratório ultrapassam a ideia de melhoria incremental para referir-se a uma nova tecnologia literária que, além de alterar o modo de comunicar o conhecimento científico, "pretende fomentar a produção aberta de conhecimento, reestruturar o processo de avaliação por pares, melhorar a qualidade da informação circulante e ampliar a participação da ciência".

Nesse sentido, esta seção apresenta as tecnologias (material, social e literária) na constituição de uma nova maneira de produzir o conhecimento. Descreve projetos com a iniciativa de publicação de dados registrados em cadernos de pesquisa, destacando as abordagens tecnológicas para publicação de dados de pesquisa adotadas nos projetos *UsefulChem*, *LabScribbles* e *Openlabnotebooks*, enquanto cadernos abertos de pesquisa de maior destaque pela literatura.

5.3.1 Tecnologias Material, Social e Literária

No movimento de compreender como os modos de ciência emergentes – a Ciência Aberta e a Ciência Comum - pretendem legitimar novas maneiras de produzir conhecimento e disputar a própria noção de ciência, Clínio (2016) busca embasamento nas etnografias de laboratório e na abordagem das três tecnologias (material, social e literária) de Shapin e Shaffer (1985)[20] para explicar como os filósofos naturais do século XVII construíram a noção de fato científico (*matter of fact*) como uma "variedade de conhecimento" tão sólida que se tornou sinônimo de ciência. Para Shapin e Shaffer (1985), a consolidação do que, na época, representava uma nova cultura epistêmica se deu pela articulação dessas três tecnologias. Segundo Knorr-Cetina (1999, p. 1), "culturas epistêmicas são culturas que criam e justificam o conhecimento, e a principal instituição do conhecimento em todo o mundo, ainda é a Ciência". Foi nessa linha de raciocínio que Shapin e Shaffer (1985) buscaram na junção das tecnologias material, social e literária uma nova forma de gerar conhecimento.

No entendimento de Shapin e Shaffer (1985 *apud* CLINIO, 2016) a tecnologia material está relacionada ao espaço do laboratório e seus instrumentos científicos rudimentares para realização de experimentos. A exemplo do laboratório de Bradley, que desenvolve o projeto *UsefulChem*, a tecnologia material está relacionada ao espaço físico e digital para o compartilhamento aberto e remoto dos agentes antimaláricos. O laboratório físico é prioritariamente de uso de membros do laboratório, mas não exclusivo. Enquanto que

[20] SHAPIN, Steven; SCHAFFER, Simon. **Leviathan and the air-pump**: Hobbes, Boyle and the experimental life. Princeton University Press, 1985.

o laboratório aberto e remoto pode ter participação de especialistas e não especialistas, possibilitando a ampliação das colaborações na ciência.

A tecnologia social refere-se a quem pode produzir o conhecimento, como os cientistas devem lidar com as controvérsias e validar o conhecimento. Nesse caso, segundo Clinio e Albagli (2017), adota-se a prática do testemunho por três tipos de testemunhas:

> (1) as testemunhas modestas dos filósofos experimentais que realizavam experimentos e apresentavam as leis por detrás dos fenômenos naturais; (2) as testemunhas diretas da seleta plateia de nobres ingleses convidados a assistir presencialmente a execução de experimentos em laboratórios e cuja participação era capitalizada para inferir credibilidade àquilo que foi testemunhado; (3) as testemunhas diretas, porém distantes ou testemunhas virtuais que participavam indiretamente dos experimentos através da leitura dos ensaios experimentais. (CLINIO; ALBAGLI, 2017, p. 8).

A tecnologia literária surge como novo tipo de arquivo que comunica o fenômeno produzido, com auxílio de instrumentos científicos para pessoas que não testemunharam diretamente o experimento. Nas palavras de Clinio e Albagli (2017, p. 8) "trata-se do gênero literário dos ensaios experimentais (*experimental essays*) que prescreve uma série de recomendações estilísticas com o objetivo de projetar na mente do leitor a sensação de simultaneidade e veracidade do experimento que está sendo narrado".

A partir da compreensão das tecnologias material, social e literária definidas por Shapin e Shaffer (1985), Clínio (2016) adota a perspectiva das três tecnologias para descrever como a tecnologia literária do caderno de laboratório se articula com novas tecnologias materiais e sociais para engendrar uma nova cultura epistêmica. A nova cultura epistêmica resultante da articulação entre as novas tecnologias é chamada de questão de prova (*matter of proof*) para valorizar a habilidade dos cientistas em documentar adequadamente os experimentos que subsidiam suas afirmações. Além disso, Clinio (2016) analisa as tecnologias em torno do laboratório de Jean-Claude Bradley para demonstrar o favorecimento de um novo modo de ciência que privilegia o caderno aberto de pesquisa.

Sendo assim, as tecnologias material, social e literária foram descritas por Clinio (2016), a partir de *posts* que registram as atividades cotidianas do projeto *UsefulChem*, para demonstrar como as três tecnologias estão embricadas em favor de um novo modo de fazer ciência, a qual é fundamentada na transparência e na proveniência dos dados.

5.3.1.1 Tecnologia Material

A tecnologia material é composta pelos equipamentos físicos e insumos químicos do laboratório de pesquisa utilizados para a realização dos experimentos. Conforme Clínio (2016,

p. 112) “o laboratório de Bradley era equipado com instrumentos científicos, computadores, substâncias químicas e insumos necessários para a realização dos experimentos”. Nesse laboratório surgiu a proposta do *Open Notebook Science* e com isso os equipamentos foram adotados também para a disponibilização dos experimentos e informações das pesquisas em ambientes digitais. Segundo Bradley (2006, não paginado) “o caderno de laboratório não precisa se parecer com um caderno de papel, mas é essencial que as informações necessárias para que o pesquisador chegue às suas conclusões estejam igualmente disponíveis para o resto do mundo”.

Figura 06 - Laboratório de Jean-Claude Bradley

Fonte: Universidade de Drexel (2014); Clinio (2016).

Clinio (2016) destaca que a tecnologia material do laboratório de Bradley colaborou ainda para a produção de conteúdos para fins acadêmicos e captura de imagens de experimentos. Ainda segundo Clinio (2016), o acesso ao espaço físico do laboratório era restrito, mas acessível por não membros da equipe. Porém, os dados das pesquisas estavam acessíveis e diversos pesquisadores puderam colaborar remotamente, via Web, surgindo parcerias e projetos relevantes. Clinio e Albagli (2017) destacam que o primeiro projeto realizado em parceria remota, o *Open Science Loop*, na qual a elaboração de hipóteses, acoplagem molecular (*docking*), síntese e os resultados da análise foram realizados de modo aberto.

Nesse contexto, Clinio (2016) percebe que o laboratório de Bradley agrega uma comunidade cientifica composta por professores e alunos de pós-graduação, bem como por *expertises* complementares de amadores, estudantes e técnicos. Além destes, existem organizações sem fins lucrativos e empresas comerciais que eventualmente colaboram com a

promoção do conhecimento aberto. Assim, Clinio e Albagli (2017) ressaltam que o acesso aberto e remoto ampliou as parcerias do laboratório e possibilitou o compartilhamento de tecnologias materiais como amostras de insumos e uso de instrumentos científicos, expandindo a pesquisa aberta sobre agentes antimaláricos. A partir desse ecossistema de colaboração, o laboratório de Bradley passou a compor uma rede de laboratórios que colaboram em prol da cura da malária.

O acesso remoto e aberto das tecnologias materiais do laboratório de Bradley diminuíam barreiras geográficas, bem como a restrição cientifica, na medida em que são disponibilizados experimentos e seus protocolos em uma linguagem mais acessível, via tecnologias da Web 2.0, o que possibilitou a colaboração de pesquisadores e não pesquisadores interessados no assunto. Portanto, a tecnologia material se articula com uma nova tecnologia social ao não exigir certificações prévias de seus colaboradores, favorecendo maior participação na ciência.

5.3.1.2 Tecnologia Social

Na perspectiva da tecnologia social, as colaborações são facilitadas na medida em que os experimentos científicos são disponibilizados em ambientes digitais de maneira que, tanto um pesquisador em início de carreira ou especialistas e não especialistas de qualquer instituição possam acessar e colaborar. Nesse sentido, Bradley (2006) menciona não se importar que as contribuições provenham de participantes laureado com um Nobel, um precoce de 14 anos ou um *bot*. Na proposta de Bradley (2006), a diminuição de barreiras de entrada pela não exigência de certificações formais confere a qualquer pessoa a capacidade de realizar contribuições relevantes, na condição de par.

A visão de Bradley (2006) apresentada por Clinio (2016) e Clinio e Albagli (2017) sobre a não exigência de certificações formais é compartilhada por Schapira e Harding (2019) ao mencionarem que o caderno aberto de pesquisa pode ser idealizado de forma clara e concisa por pesquisadores iniciantes para mostrarem suas habilidades técnicas e sua visão científica, além de se conectarem com seus colegas e com especialistas na área, iniciar novas colaborações e construir sua própria rede. Nesse cenário, os pesquisadores podem ter seus próprios cadernos abertos e prestar suas colaborações aos experimentos de colegas independentes de titulações acadêmicas.

Outra vertente trazida com a perspectiva da tecnologia social é a mudança do formato de comunicação da informação entre pesquisadores, deixando de ser exclusivamente via artigos científicos, a partir de resultados finais, para se tornar o intercâmbio de resultados

parciais. Para Clinio (2016), esse intercâmbio se diversifica quantitativamente e qualitativamente. Nessa perspectiva, Clinio e Albagli (2017) citam o formato *storyless experiments* para demonstrar que o caderno aberto atribui relevância para todas as etapas de uma pesquisa, independente de seu status – em andamento, finalizado, abandonado, favorável ou ambíguo - e exime o pesquisador do papel de autor de narrativas coerentes sobre sua pesquisa. "Já a adoção de licenças abertas permite que os pesquisadores sejam autores de contribuições sem sua conversão em propriedades de informação" (CLINIO; ALBAGLI, 2017, p. 12). Esse modelo de fazer ciência valoriza as várias formas de autoria e registra as colaborações, quando, onde e como de maneira independente (CLINIO, 2016, CLINIO; ALBAGLI, 2017).

Essas colaborações podem partir de experiências em diferentes eixos dentro de uma pesquisa, o que possibilita o questionamento de afirmações e omissões ou um simples alerta de que algo não está indo bem, estabelecendo um processo permanente de revisão por pares, no qual potenciais colaboradores acessam a documentação detalhada de experimentos e podem verificar sua adequação ao tentar obter o mesmo conjunto de evidências que o autor original afirma ter produzido. Os participantes atuam como testadores em um novo modo de comunicação que não se baseia mais na ideia de 'fontes confiáveis', mas promove a transparência e a proveniência de dados, pois considera que o acesso a todo o conjunto de dados proporciona a reanálise dos dados, por isso é mais valiosa que a análise de subconjuntos apresentados de maneira condensada nos *papers* como resultados finais (CLINIO; ALBAGLI, 2017).

5.3.1.3 Tecnologia Literária

A tecnologia literária, no contexto de cadernos abertos, faz parte do ecossistema de produção e publicação de dados de pesquisa, a partir do uso de tecnologias em formato aberto. Para Clinio (2016, p. 117), a tecnologia literária reflete "a produção de novos tipos de arquivos para a documentação da prática científica que implicam no estabelecimento de padrões para entrada de informação, a adoção de formatos abertos e diretrizes para a gestão compartilhada de dados".

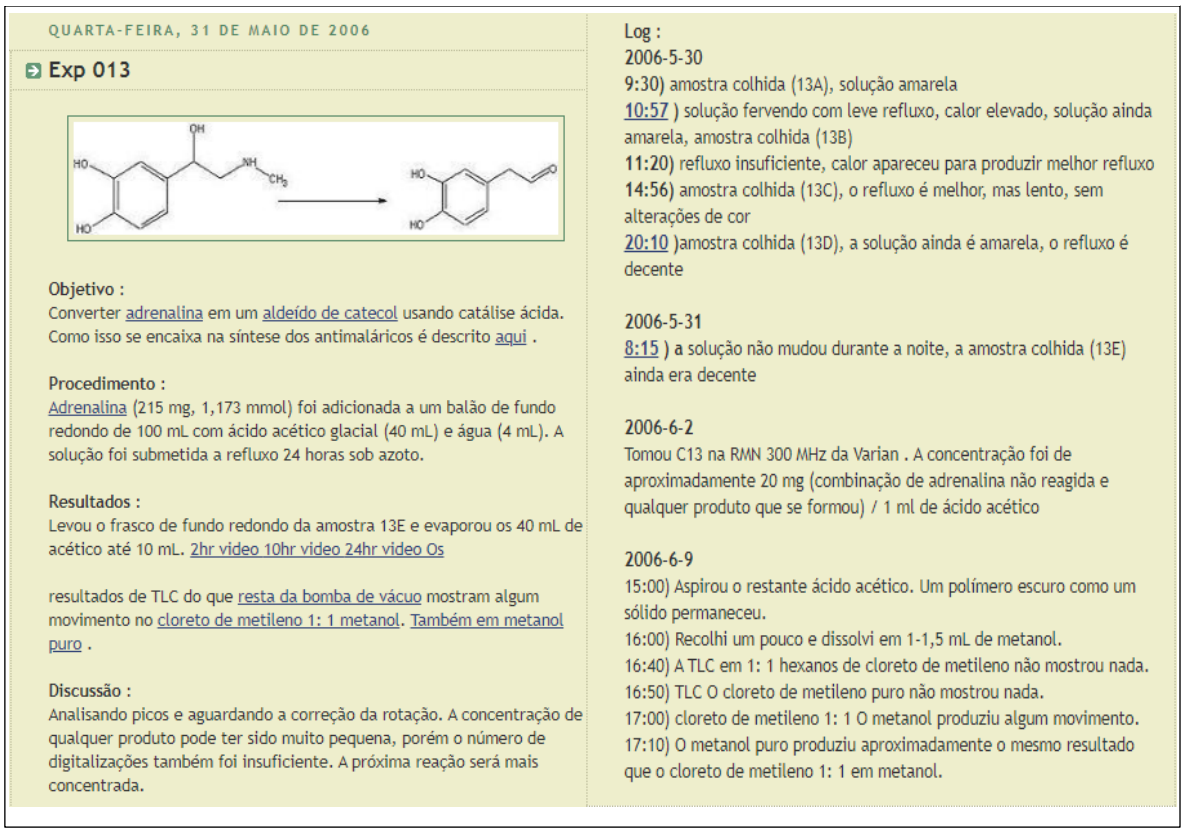
O caderno de pesquisa científica produzida e publicada em formato aberto permite a disponibilidade para o acesso, uso e reuso, em tempo real, à íntegra dos dados brutos produzidos em laboratórios. Nesse cenário, Clinio (2016) ressalta que o laboratório demanda o desenvolvimento de infraestruturas que preservam a proveniência dos dados e promovam a circulação do conhecimento científico, percorrendo desde o registro de dados brutos de

pesquisa, à descrição no formato de experimentos sem história (*storyless experiments*) até a alimentação de aplicativos e banco de dados especializados. Em se tratando de bancos de dados, Clinio (2016) reforça que os cadernos abertos devem ser plataformas amigáveis (*user friendly*) que facilitam a sua adoção por pesquisadores e integram sistemas de informação projetados para a gestão aberta do conhecimento, lidando com questões como a interoperabilidade entre plataformas; a automação de tarefas, o desenvolvimento de tecnologias da Web Semântica e aplicações.

No contexto das plataformas amigáveis, Bradley registrava seus experimentos em um formato estruturado e padronizado para garantir a qualidade do registro e possibilitar a sua reutilização em outros serviços de informação. O projeto iniciou com as postagens de detalhes de seus experimentos em um *blog,* criado na plataforma *Blogger*, pois conforme Clinio e Albagli (2017), as funcionalidades dessa plataforma atendiam às expectativas de publicar *posts*, comunicar com os colaboradores via mensagem de texto e realizar buscas. Logo, Bradley percebeu que as funcionalidades do *blog* eram incipientes para a proposta de um caderno aberto de dados de pesquisa e adota a plataforma *wiki* que apresentava as funcionalidades mais robustas para edição de conteúdos, conforme apresentado no item 5.3.3.1, sobre o projeto *UsefulChem*.

Ainda de acordo com Clinio e Albagli (2017), o caderno de Bradley foi divido em seções: (1) Número do experimento; (2) Representação gráfica do experimento; (3) Nome do pesquisador; (4) Objetivo; (5) Procedimento; (6) Resultado; (7) Discussão; (8) Conclusão e (9) Log.

Figura 07 - Caderno UsefulChem. Detalhes Experimentais



QUARTA-FEIRA, 31 DE MAIO DE 2006

Exp 013

Objetivo :
Converter adrenalina em um aldeído de catecol usando catálise ácida. Como isso se encaixa na síntese dos antimaláricos é descrito aqui .

Procedimento :
Adrenalina (215 mg, 1,173 mmol) foi adicionada a um balão de fundo redondo de 100 mL com ácido acético glacial (40 mL) e água (4 mL). A solução foi submetida a refluxo 24 horas sob azoto.

Resultados :
Levou o frasco de fundo redondo da amostra 13E e evaporou os 40 mL de acético até 10 mL. 2hr video 10hr video 24hr video Os

resultados de TLC do que resta da bomba de vácuo mostram algum movimento no cloreto de metileno 1: 1 metanol. Também em metanol puro .

Discussão :
Analisando picos e aguardando a correção da rotação. A concentração de qualquer produto pode ter sido muito pequena, porém o número de digitalizações também foi insuficiente. A próxima reação será mais concentrada.

Log :
2006-5-30
9:30) amostra colhida (13A), solução amarela
10:57) solução fervendo com leve refluxo, calor elevado, solução ainda amarela, amostra colhida (13B)
11:20) refluxo insuficiente, calor apareceu para produzir melhor refluxo
14:56) amostra colhida (13C), o refluxo é melhor, mas lento, sem alterações de cor
20:10)amostra colhida (13D), a solução ainda é amarela, o refluxo é decente

2006-5-31
8:15) a solução não mudou durante a noite, a amostra colhida (13E) ainda era decente

2006-6-2
Tomou C13 na RMN 300 MHz da Varian . A concentração foi de aproximadamente 20 mg (combinação de adrenalina não reagida e qualquer produto que se formou) / 1 ml de ácido acético

2006-6-9
15:00) Aspirou o restante ácido acético. Um polímero escuro como um sólido permaneceu.
16:00) Recolhi um pouco e dissolvi em 1-1,5 mL de metanol.
16:40) A TLC em 1: 1 hexanos de cloreto de metileno não mostrou nada.
16:50) TLC O cloreto de metileno puro não mostrou nada.
17:00) cloreto de metileno 1: 1 O metanol produziu algum movimento.
17:10) O metanol puro produziu aproximadamente o mesmo resultado que o cloreto de metileno 1: 1 em metanol.

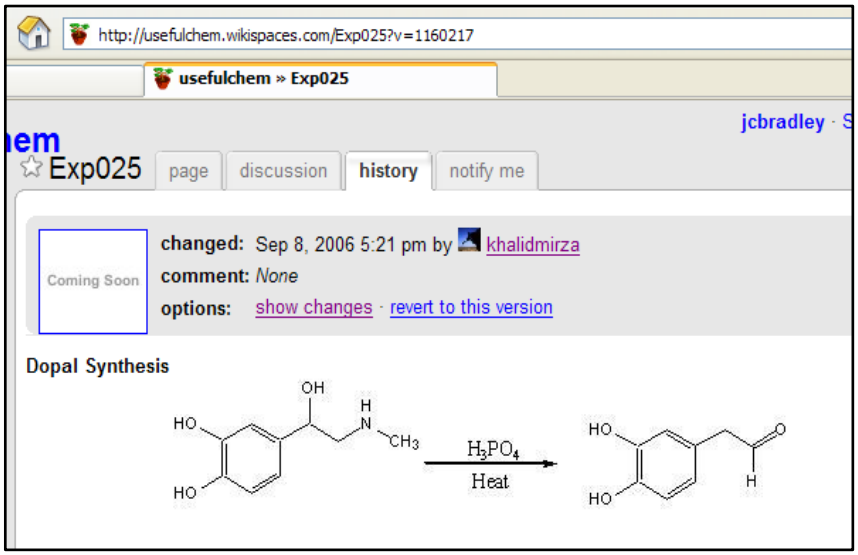
Fonte: Blog UsefulChem Experiments. Experimento 013 (2006).

O *log* é a seção que descreve cada etapa do experimento. Clinio e Albagli (2017) explicam que é a seção mais importante do caderno de Bradley, pois ao descrever o experimento enquanto está sendo realizado, garante a precisão e riqueza de detalhes.

As experiências em andamento passaram a ser publicadas na plataforma *wiki* do *UsefulChem*. Nessa plataforma, Bradley (2007) apresenta a possibilidade de acompanhar as mudanças feitas no experimento, por meio da aba ‘histórico’ do caderno de pesquisa desenvolvido na *wiki*. Na plataforma *wiki*, é possível visualizar as alterações mais recentes e comparar o que foi adicionado, bem como o que foi excluído na experiência. Segundo Bradley (2007) é possível saber não apenas as mudanças, mas também como o pesquisador, o aluno, o supervisor e os colegas chegaram aos resultados, apresentaram argumentos na discussão e chegaram às suas conclusões.

Nesta configuração de caderno aberto a revisão é aberta e os próprios pesquisadores verificam a qualidade da contribuição ao utilizá-la, além dos dados ficarem disponíveis na medida em que são gerados.

Figura 08 - Caderno UsefulChem. Experimento na plataforma Wiki



Fonte: Bradley (2007).

De acordo com Bradley (2007) essas atividades não são possíveis de serem feitas a partir do *blog*. Portanto, as postagens referentes aos experimentos do projeto *UsefulChem* passaram a ser realizados na plataforma *wiki* e os dados do *blog* ficaram registrados para efeitos históricos.

O laboratório de Bradley foi o primeiro projeto *open notebook Science*, por isso apresenta tamanha importância, no entanto atualmente as plataformas não estão disponíveis para verificação dos padrões e tecnologias adotados. As ferramentas do projeto *UsefulChem* estão descritas no item 5.3.3.1, a partir de uma revisão de literatura.

5.3.2 Caderno Aberto e a Construção de uma nova Cultura Epistêmica

O caderno aberto à luz da composição das tecnologias material, social e literária é considerado uma inovação para um novo modo de produzir ciência. Nesse cenário, o caderno aberto de pesquisa é elemento de uma nova cultura epistêmica. Essa nova cultura epistêmica se baseia, segundo Clinio (2016), em uma variedade de conhecimento denominando uma questão de prova (*matter of proof*), na qual a possibilidade de comunicar a informação satisfatória para terceiros não depende da obtenção de um resultado final favorável e aprovado pelo tradicional modelo de revisão por pares, mas fundamentada na qualidade da documentação detalhada e no registro minucioso das etapas intermediárias de um experimento como parte da sua estratégia para promover discussões mais consistentes em ciência e colaboração aberta. Esse modelo de produção e compartilhamento de dados não tem como alicerce a construção de fatos científicos (*matter of fact*) que se caracteriza pela ciência

pronta, ou seja, resultado final e satisfatório do ponto de vista daquela pesquisa, em que dados brutos não aparecem para compartilhamento e colaboração. Além disso, a noção de fato científico fomenta a cultura do segredo cujos dados são mantidos em sigilo até o resultado final da pesquisa, procrastinando o desenvolvimento da ciência e, consequentemente, de novas descobertas e soluções.

Sendo assim, um modelo semântico para estruturação e publicação de dados abertos de cadernos de pesquisa fortalecerá a proposta dessa nova cultura epistêmica.

A partir do entendimento do caderno aberto de pesquisa como elemento da nova cultura epistêmica, passa-se para a identificação de projetos com iniciativas de publicação de cadernos abertos de pesquisa, com ênfase nas tecnologias adotadas para publicação e compartilhamento, em tempo real, dos dados de pesquisa científica.

5.3.3 Projetos de Cadernos Abertos de Pesquisa

Projetos com a iniciativa de publicação de dados de pesquisa com o uso de tecnologias abertas foram criados por pesquisadores com o intuito de promover a transparência, a colaboração entre pesquisadores e descobertas mais rápidas. Dentre as iniciativas que adotam tecnologias abertas identificadas como ativas e inativas incluem:

Open Source Malaria (OSM)[21], coordenado por Matthew Todd e busca novos medicamentos para a cura da malária. O projeto OSM tem como princípios compartilhar todos os dados e ideias abertamente, com a permissão para qualquer pessoa contribuir e não há patentes. O OSM recebe de pesquisadores contribuições em projetos existentes e abre espaço para novos projetos.

Open Notebook Science Network (ONSN)[22], liderada por Brian Glanz e com orientações dos pesquisadores de cadernos abertos Anthony Salvagno e do falecido Jean-Claude Bradley. A rede ONSN apoia pesquisadores de várias áreas. A ferramenta é construída em *WordPress* e possui licença de atribuição *Creative Commons*.

O laboratório de Rosanne Hertzberger, identificado na Web pelo nome RebLab[23], estuda o metabolismo dos micróbios vaginais. O projeto tem um caráter completamente aberto e tem como propósito aumentar a eficiência cientifica, compartilhando o máximo de informações com outros pesquisadores e o público geral.

---

[21] *Open Source Malaria* (OSM). Disponível em: http://opensourcemalaria.org
[22] *Open Notebook Science Network* (ONSN). Disponível em: http://onsnetwork.org Catalog
[23] RebLab. Disponível em: http://www.reblab.org

Estas iniciativas publicam os experimentos em *blogs* e em repositórios de dados aberto como o Zenodo.

Outros projetos de grande destaque na literatura são *UsefulChem*, primeira iniciativa de caderno aberto desenvolvido no laboratório de Bradley; *LabScribbles* de Rachel Harding para compartilhar dados sobre a doença de *Huntington*; e *Openlabnotebooks* desenvolvido pelo Structural Genomics Consortium (SGC) a partir da experiência do *LabScribbles*, os quais serão descritos nos itens seguintes.

5.3.3.1 Projeto UsefulChem

O projeto *Useful Chemistry* ou *UsefulChem* foi iniciado no laboratório de Bradley, em 2005, como uma iniciativa baseada na Web para tornar o processo científico o mais transparente possível, publicando todo o trabalho de pesquisa em tempo quase real para uma coleção de *blogs* públicos, *wikis* e outras páginas da Web.

Para Bradley (2010), o *UsefulChem* é uma proposta *Open Notebook Science* de química orgânica, com o objetivo principal de descobrir novos agentes antimaláricos que pudessem ser preparados por sínteses simples e baratas. Isso representa uma oportunidade para todos se beneficiarem das pesquisas em andamento. No entanto, para fazer uso da informação, ela deve ser facilmente encontrável. Uma estratégia simples para aumentar a capacidade de descoberta seria disponibilizar todo o conteúdo registrado no caderno de laboratório, bem como os arquivos de dados brutos associados, em várias plataformas de comunicação digital. Assim, a trajetória do projeto *UsefulChem* se desenvolveu a partir do uso de diversas ferramentas digitais.

As primeiras iniciativas de caderno aberto de laboratório começaram com a criação do blog *UsefulChem* hospedado no Blogger, serviço oferecido pelo Google, para descobrir e trabalhar com temas urgentes de química, registrar as atividades de pesquisa e relatar o progresso do projeto de maneira transparente.

O *blog UsefulChem* atuou como o projeto maior, guarda-chuva, e foi se desmembrando em outros *blogs* com temas e atividades específicas, como o *UsefulChem Experiments* e *UsefulChem Molecules*. Dessa forma, destaca-se que os primeiros experimentos de laboratório foram registrados no novo *blog UsefulChem Experiments* e informações sobre moléculas relevantes foram coletadas em *posts* no blog *UsefulChem Molecules*. Bradley, Lang, Koch e Neylon (2011) lembram que o pesquisador Mathew Todd, da Universidade de Sydney, estreou o *blog UsefulChem Experiments* com comentários

enriquecedores que resultaram em um novo projeto aberto do Grupo Todd, sobre o produto químico praziquantel.

No inicio, a ferramenta atendia as expectativas de atualização rápida a partir de acréscimos de postagens e permitia o diálogo entre os colaboradores, via caixa de comentários, bem como o recurso de busca e RSS *feed* para os leitores. Segundo Clinio e Albagli (2017), as funcionalidades do *blog* eram incipientes, pois a colaboração aberta exigia mais do que comentar, sugerir e criticar. Para as autoras era preciso editar diretamente os conteúdos, reescrever, refazer sem a necessidade de solicitar um *login* e senha ao administrador da plataforma. Assim, foi criado um *wiki* para organizar informações coletivas, vinculando-se a publicações relevantes ou outros recursos. Para Cameron (2008), uma das razões de adotar o *wiki* é a possibilidade de melhorar a descrição dos experimentos com o registro de rastreamento de versão apropriado para determinar quem contribuiu com o que e quando. Nesse sentido, Bradley, Lang, Koch e Neylon (2011), defendem o uso de *wiki* para um caderno de laboratório também se tornou conveniente para os mentores se comunicarem com os alunos, comentando diretamente em seções específicas de uma página. A disponibilidade de alertas por e-mail para quaisquer alterações no *wiki* facilitou uma comunicação mais rápida. Ainda segundo Bradley, Lang, Koch e Neylon (2011), o *Wikispaces* foi a plataforma escolhida para o projeto, pois oferece um serviço gratuito hospedado para *wikis* públicos e oferece editor visual intuitivo, *wikitext* simplificado e recursos convenientes de *backup* e alerta.

Com o aumento de dados gerados no laboratório, esforços foram realizados para o fornecimento de ferramentas de busca para a recuperação da informação. Assim, o projeto adotou o serviço *Google Co-op Search*, o atual Pesquisa Personalizada do Google, que permite realizar uma pesquisa federada de todas as plataformas *UsefulChem*. Para o armazenamento e manipulação de dados, foram adotadas as planilhas do Google. Em março de 2007, o projeto testou a base de dados de substâncias químicas *ChemSpider* para gerenciar as moléculas geradas pelo *UsefulChem*. A transição completa para o *ChemSpider* ocorreu em julho de 2007, com a demonstração da busca de subestruturas e o *blog UsefulChem-Molecules* foi descontinuado. Em agosto de 2007, a Collaborative Drug Discovery (CDD), empresa privada de solução para banco de dados para descoberta de drogas, forneceu uma conta gratuita para armazenar e compartilhar os resultados dos ensaios produzidos pela *UsefulChem*. Em abril de 2008, o *UsefulChem* adquiriu um subdomínio no *ChemSpider*. Outra ferramenta adotada pelo projeto foi a *FriendFeed* para a agregação de *feeds*.

Dentre os projetos de Bradley encontra-se o Repositório *UsefulChem* – hospedado no re3data –, o qual apresenta característica *Open Notebook Science* em química, liderado pelo *Bradley Laboratory* na Drexel University. O re3data descreve que o principal projeto do *UsefulChem* envolve a síntese de novos compostos antimaláricos e apresenta descrições detalhadas dos experimentos. O *UsefulChem* teve entrada no re3data em abril de 2014 e a sua última atualização realizada em fevereiro de 2019.

5.3.3.2 Projeto LabScribbles

O *LabScribbles*, rabiscos de um laboratório, é uma iniciativa de caderno aberto de pesquisa desenvolvida por Rachel Harding, em 2016, durante o pós-doutorado em *Structural Genomics Consortium* (SGC), na Universidade de Toronto. Harding (2016) compartilha seus dados experimentais brutos, em tempo real, sobre a doença de Huntington – doença neurodegenerativa hereditária – para que outros cientistas neste campo possam avaliar e colaborar com estudos sobre a doença.

Para Harding (2016), a ciência é muito secreta e lenta. A autora explica que o gene que causa a doença de Huntington foi identificado em 1993, mas nas décadas seguintes, os cientistas não descobriram como exatamente esses genes causam os sintomas associados aos distúrbios. Nas palavras de Harding (2016) "sabemos que, de alguma forma, [o gene] distorce a proteína que está codificada e altera suas funções". Sem saber como essas proteínas funcionam, a cura para a doença de Huntington está muito distante. Sendo assim, a exemplo de Bradley, a pesquisadora está compartilhando todos os seus dados em mídias sociais. Através desses compartilhamentos, espera-se criar um *ethos* colaborativo com outros cientistas com o propósito de acelerar a taxa de entregas de dados que possam informar oportunidades terapêuticas potenciais.

No entendimento de Harding (2019) a ciência pode acelerar descobertas se adotados os princípios da ciência honesta em tempo real, acessível, interativa e de acesso aberto. Por isso, apresenta seus dados brutos diferentes das publicações finais, mas uma amostra da vida real, do funcionamento diário e da realidade de um cientista. O *LabScribbles* foi projetado para ser acessível, claro e detalhado em sua apresentação, para permitir o diálogo entre leitores e abrir o caminho para colaborações. Assim, as postagens incluem um resumo de cada publicação explicando cada experimento, os métodos, seus resultados e a relevância geral sem muito jargão científico. A análise desses dados e as publicações relevantes também são postadas no *blog*.

Na pesquisa em andamento Harding (2019) faz um resumo do experimento, com uma linguagem acessível, e publica no *blog LabScribbles*, ligando o resumo ao conjunto de dados, com o objetivo de alcançar um público mais amplo e não especializado. Nesse sentido, o objetivo de Harding (2019) é contextualizar os experimentos dentro do projeto de pesquisa mais amplo, além de fornecer orientações sobre pontos de melhoria.

Conforme Harding (2019), os dados são publicados em formatos de arquivo adequados e os carrega para o repositório de acesso aberto Zenodo, como um conjunto de dados licenciado com o *Creative Commons Attribution.*

A pesquisadora Harding (2019) espera que a transparência incentive outras pessoas a colaborar em seus projetos por meio de correções e a contribuir para elevar a ciência à velocidade da era da Internet. Portanto, acredita-se que os *blogs* possam aumentar a transparência em torno da pesquisa, mesmo sendo precoce afirmar que seja o melhor formato para incentivar o progresso e a colaboração cientifica.

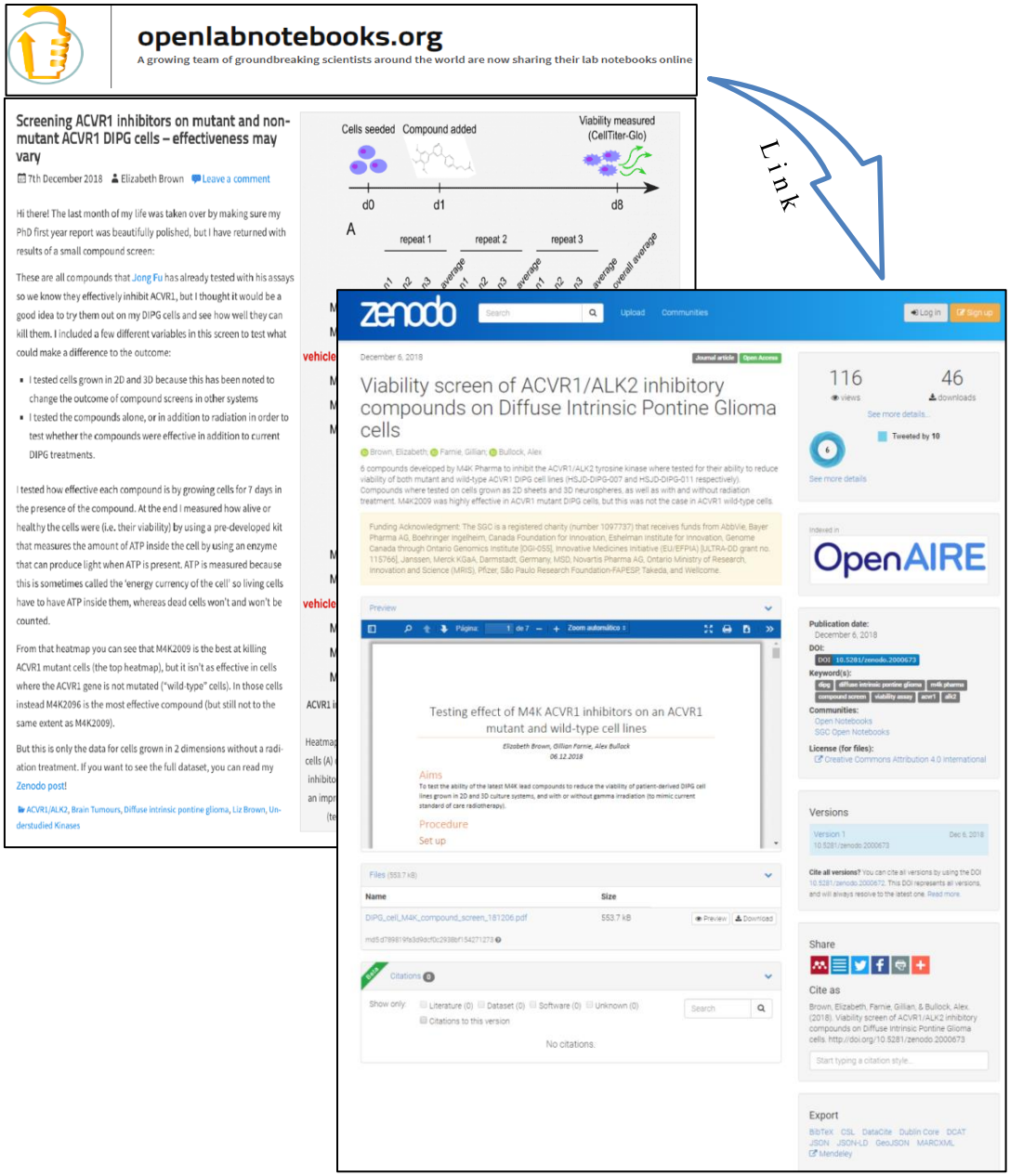
O formato de caderno adotado por Harding (2019) foi inspirado no trabalho de outros cientistas que publicam dados abertos e influenciou também a criação do projeto *Openlabnotebooks*, o qual segue as mesmas diretrizes e formatos de publicação.

5.3.3.3 Projeto Openlabnotebooks

O projeto *Openlabnotebooks* é uma iniciativa de caderno aberto de pesquisa, lançado em janeiro de 2018, a partir da experiência do *LabScribbles*, pelo Structural Genomics Consortium (SGC) - organização que apoia a pesquisa de proteínas medicamente importantes -, para compartilhar pesquisas científicas em andamento com o propósito de gerar novas ideias e discussões, evitar redundâncias e acelerar descobertas. De acordo com Schapira e Harding (2019), o lançamento do *blog Openlabnotebooks* contou com a participação inicial de 12 (doze) pesquisadores do SCG relatando seus trabalhos ao vivo e on-line.

Segundo Schapira e Harding (2019), a participação no *Openlabnotebooks* é composta por duas postagens, sendo (1) no *blog*, os dados experimentais são dirigidos para não especialistas com informações sobre a motivação de realização do experimento, o resumo e os resultados alcançados, como também os delineamentos dos próximos passos e indica o *link* para detalhes dos experimentos; (2) no repositório Zenodo publica-se o registro experimental detalhado e rigoroso, incluindo todos os dados e protocolos, os quais os especialistas possam avaliar, comentar ou construir. Na figura 09 é possível visualizar a publicação no *blog Openlabnotebook* e no repositório Zenodo.

Figura 09 - Publicação de dados *no blog Openlabnotebook* e no repositório Zenodo



Fonte: Elaborado a partir de *LabScribbles* (2019).

O pesquisador que busca informação no *blog Openlabnotebooks* visualizará a opção de acessar os detalhes dos dados por meio do *link* de acesso ao repositório Zenodo, no final de cada postagem.

Dentre as ferramentas tecnológicas adotadas pelo *Openlabnotebook*, como pode ser visto na figura 09, incluem os *blogs* que são gerenciados por um servidor baixado do *wordpress.org* e vinculados aos registros experimentais, que são depositados no repositório de dados abertos Zenodo. De acordo com Schapira e Harding (2019), os dados experimentais disponibilizados no *blog Openlabnotebooks* podem também ser gerenciados por outros repositórios públicos como GitHub[24] ou Figshare. Algumas tecnologias legíveis por máquinas estão presentes no repositório Zenodo, como CSL, Dublin Core, JSON e MARCXML.

Para os autores, embora os detalhes experimentais postados no Zenodo sejam importantes cientificamente, o *blog* escrito em termos leigos pode ser usado para envolver pesquisadores que possam ter um conjunto complementar de conhecimentos para futuras colaborações, bem como outras partes interessadas no processo de pesquisa, incluindo grupos de pacientes.

5.4 ESTRUTURA DOS DADOS DE PESQUISA CIENTÍTICA DE CADERNOS DE PESQUISA

Para apresentar a estrutura dos dados de pesquisa científica publicados pelos cadernos abertos de pesquisa, buscou: 1- na literatura, identificar a tipologia e as especificidades dos dados; 2 - nos cadernos *UsefulChem*, *LabScribbles* e *Openlabnotebooks*, identificar tais características na prática; e 3 – Comparar as práticas adotadas entre os cadernos *UsefulChem*, *LabScribbles* e *Openlabnotebooks* quanto aos formatos, padrões e estrutura de dados.

5.4.1 Tipologia de Dados de Pesquisa Científica de Cadernos de Pesquisa

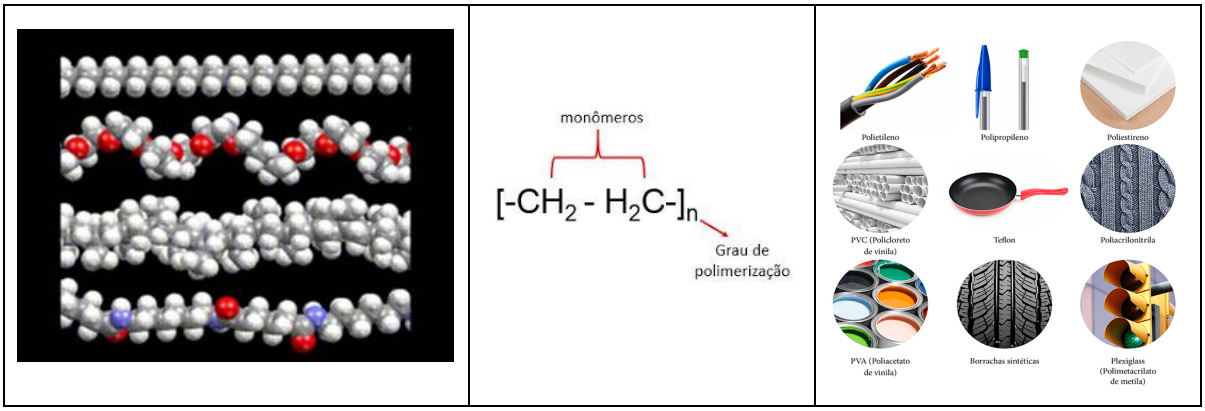
O principal aspecto que diferencia os dados registrados em cadernos de pesquisa dos demais tipos de dados de pesquisa científica está em suas características e tipologias, os quais são explicitados nas descrições de metadados. De acordo com o National Science Board (NSB) (2005), os dados gerados em laboratórios são classificados como dados experimentais. Para Silva (2019, p. 28) dados experimentais "são resultados de procedimentos realizados em condições controladas com a finalidade de provar o estabelecimento de hipótese sobre determinado fenômeno". Para Borgman (2010), os dados experimentais incluem resultados de

[24] Github. Disponível em: github.com

estudos de laboratório, como medições de reações químicas, ou de experimentos de campo, como estudos comportamentais controlados.

As medições de reações químicas são um procedimento frequente nos laboratórios de pesquisa. Sendo assim, consultaram-se informações sobre possíveis pontos de acesso para buscas em uma plataforma digital, de maneira a definir metadados adequados com vista a recuperação e uso. Segundo Mettler Toleto (2020) especialistas garantem que reações bem reguladas produzem moléculas que são caracterizadas com relação à composição, peso molecular, distribuição de peso molecular e propriedades físicas estruturais. Um exemplo de resultado de reação química presente no cotidiano humano encontra-se nos polímeros sintéticos, como náilon e o poliuretano, que transformaram a manufatura e o uso de produtos comerciais. A título de compreensão, os polímeros são macromoléculas formadas a partir de unidades estruturais menores, ou ainda, são macromoléculas que consistem de subsegmentos monoméricos repetidos ligados para formar cadeias.

Figura 10 - Modelo de Reação Experimental



Fonte: Google Imagens (2020).

Segundo Mettler Toleto (2020), os gráficos devem ser apresentados e usados em conexão com a descrição sobre a pesquisa para que não haja interpretação errônea ou tendenciosa, afetadas pelas concepções teóricas de cada pesquisador.

Sendo assim, é recomendável que os pesquisadores apresentem a descrição mais completa possível, incluindo o registro de datas e horário de realização e modificação dos experimentos, nomes dos reagentes, objetivos pretendidos, resultados alcançados, dados referentes ao pesquisador, do laboratório e da instituição vinculada, nível de abertura dos dados e utilização de formatos de armazenamento adequados. Essas informações são fundamentais para a curadoria e preservação dos dados.

De acordo com Gold (2007) a diferenciação dos tipos de dados segundo a sua origem ou coleta são cruciais para a decisão da melhor forma de armazenamento e preservação dos dados. Para o NSB (2005) e Gold (2007) os dados experimentais podem não ser facilmente reproduzidos, considerando os custos e a complexidade dos experimentos. Os fatores custo e reprodutividade são relevantes para a definição de politicas de preservação de dados experimentais, pois quando os dados forem processados para uma variedade de finalidades, uma infinidade de produtos derivados deverá ser preservada.

O NSB (2005) sugere ainda que os dados possam ser diferenciados com base em sua natureza, por exemplo, números, imagens, fluxos de vídeo ou áudio, informações sobre a versão do software, algoritmos, equações, animações ou simulações. Os dados podem ainda ser classificados conforme a sua fase de pesquisa, sendo: dados brutos (*raw data*), dados derivados e dados canônicos ou referenciais. A classificação segundo a natureza dos dados pode compreender parte dos elementos de um experimento, os quais podem facilitar a compreensão dos dados experimentais. Da mesma forma, dados experimentais podem ser combinados e gerados novos dados.

Entende-se que os dados de pesquisa científica registrados em cadernos de pesquisa sejam comumente de origem experimental, considerado a sua classificação de coleta, porém Silva (2019) ressalta que vários tipos de registros estão associados com os dados observacionais, experimentais e computacionais, tais como os registros históricos, os registros de campo ou as notas manuscritas, bem como a gravação de áudio que raras vezes se define, apesar de ter extenso uso na linguagem diária. Nesse caso, Silva (2019) explica que pode ser considerada como uma nova categoria de dados, tal qual a sua tipologia de coleta, pois não se define em nenhuma das citadas na literatura.

Os filtros de busca definidos pelos cadernos *LabScribbles* e *Openlabnotebooks* indicam como tipos de dados o conjunto de dados, publicação, pôster e apresentação.

As definições de Sayão e Sales (2015, p. 79) para o termo geral ‘conjuntos de dados ou *datasets’* é usado para descrever uma coleção de dados de pesquisa “que pode ser formado por um único elemento, como uma planilha de dados numéricos; pode igualmente ser formado por um conjunto de elementos relacionados, tais como planilhas, imagens, ou leituras diárias de um instrumento científico”. Os arquivos recuperados nos cadernos *LabScribbles* e *Openlabnotebooks* como conjunto de dados são, com frequência, o mesmo documento em diferentes tipos de extensão de arquivo, como .docx, .xlsx e .pdf.

O tipo de dados denominado de ‘publicações’ traz documentos como artigos, relatórios, pré-impressão (*preprint*), *paper* de conferência, documento de trabalho,

frequentemente em arquivos de extensão docx, xlsx e pdf. O ‘pôster’ é o documento organizado para apresentação em congresso científico, geralmente contém apenas uma página com informações sucintas e com imagens, no formato pdf. O tipo de documento chamado de ‘apresentação’ traz informações organizadas em *slides* para fins de palestras, aulas e treinamentos, normalmente o documento aparece com a extensão ppt e pdf.

Os dados de pesquisa que o caderno *UsefulChem* publicava era constituído por aulas e palestras armazenadas em arquivos .ppt, representações gráficas em formatos JPEG, áudios e textos armazenados em arquivos de extensão pdf.

Os tipos de dados publicados em cadernos de pesquisa de laboratório podem ser visualizados no quadro 12, a seguir

Quadro 12 - Tipos de Dados

Tipos de Dados		**Descrição**
Origem	Experimental	Resultados de procedimentos realizados em condições controladas com a finalidade de provar o estabelecimento de hipótese sobre determinado fenômeno.
Natureza	Textos	Manifestação linguística das ideias de um autor, que serão interpretadas pelo leitor de acordo com seus conhecimentos linguísticos e culturais.
	Tabelas	Quadro sistemático de consulta de dados.
	Gráficos	Curva num sistema de coordenadas, que representa uma função .
	Números	Descreve quantidades, ordem, medida.
	Fórmulas	Expressão concisa e rigorosa, constituída em geral de símbolos, que resumem certo número de dados.
	Equações	Redução de uma questão, um problema intrincado, a pontos simples e claros, para facilitar a obtenção de uma solução.
	Algoritmos	Sequência finita de regras, raciocínios ou operações que, aplicada a um número finito de dados, permite solucionar classes semelhantes de problemas.
	Imagens	Representação da forma ou do aspecto de ser ou objeto por meios artísticos.
	Fotografias	Técnica de criação de imagens por meio de exposição luminosa, fixando-as em uma superfície sensível.
	Áudios	Técnica de transmissão, recepção e reprodução de sons.
	Vídeos	Técnica de reprodução eletrônica de imagens em movimento.
	Filmes	Sequência de imagens registradas em filme cinematográfico ou videoteipe, para exibição em movimento ou não.
	Preprint	A versão de um manuscrito antes da avaliação por pares, os quais certificam ou não sua publicação formal em um periódico.
	Resultados de experimentos	Resultado de trabalho científico que se destina a verificar um fenômeno.
	Bases de dados	Conjunto de dados inter-relacionados sobre determinado assunto, armazenados em sistemas de processamento de dados segundo critérios preestabelecidos (reúne).
	Simulações	Teste, experiência ou ensaio em que se empregam modelos para simular o ser humano, em especial em casos de grande perigo de vida.
Fases	Derivados	Resultados de combinação de dados brutos ou de outros dados.
	Brutos	São dados que vêm diretamente dos instrumentos científicos.
	Canônicos ou referenciais	São coleções de dados consolidados e arquivados geralmente em grandes centros de dados, por exemplo, sequência genética, estrutura química etc.

Fonte: Adaptado de Sales (2014).

A partir da identificação dos tipos e características dos dados registrados em cadernos de pesquisa passa-se para a análise dos formatos apresentados pelos cadernos em estudo.

5.4.2 Formatos dos Dados de Pesquisa Científica dos Cadernos de Pesquisa

Em consulta às plataformas dos projetos *UsefulChem*, *LabScribbles* e *Openlabnotebooks*, buscou-se identificar os tipos de dados que os cadernos armazenam, os tipos de metadados adotados para descrever os dados, formatos de dados, vocabulários, grau de estruturação, licença de uso e a plataforma digital adotada para publicar os dados.

Quadro 13 – Elementos estruturais dos cadernos estudados

Estrutura	**Cadernos Abertos de Pesquisa**		
	Openlabnotebooks	***LabScribbles***	***Openlabnotebooks***
Tipos de dados Publicados	vídeos, aulas, palestras, entrevistas capítulos de livros e artigos	Conjunto de dados, publicação, pôster e apresentação	Conjunto de dados, publicação, pôster e apresentação
Metadados Descritivos	Essenciais	Mínimo	Mínimo
Metadados Proveniência	Não identificado	Não identificado	Não identificado
Metadados Preservação	Não identificado	Não identificado	Não identificado
Vocabulários	Não identificado	JSON, JSON-LD, *RDF/XML*, DCAT, MARCXML, BibTex, OGP	JSON, JSON-LD, *RDF/XML*, DCAT, MARCXML, BibTex, OGP
Tipo de documento	pdf e ppt	pdf, docx, xlsx e rdf	pdf, docx, xlsx e rdf
Outros tipos	JPEG	Não identificado	Não identificado
Grau de estruturação	Não estruturado	Semiestruturados	Semiestruturados
Licença de uso	Não identificada	Creative Commons Attribution 4.0	Creative Commons Attribution 4.0
Plataforma	Blog	Zenodo e blog	Zenodo e blog

Fonte: Elaborado pela autora (2020).

Dentre os dados armazenados nos cadernos *UsefulChem*, *LabScribbles* e *Openlabnotebooks*, apresentados no quadro 13, estão textos, tabelas, gráficos, números, fórmulas, representação gráfica do experimento, traços, mapas, fotografias e áudios gravados em diferentes tipos de extensão de dados, inseridos nos objetos nomeados de conjunto de dados, publicação, pôster e apresentação, conforme definições apresentadas no item 5.4.1.

O caderno *UsefulChem* fornece metadados descritivos legíveis por humanos e não por máquinas. Os atributos dos dados publicados nesse caderno, figuras 7 e 8, são: o número do experimento, ou seja, cada experimento realizado possui um código como se fosse um tombamento do experimento; pode conter a representação gráfica do experimento; nome do pesquisador; objetivo da pesquisa; procedimentos realizados, resultados alcançados;

discussão; conclusão; e *log*. O *log* é a descrição do experimento enquanto está sendo realizado para garantir a precisão e riqueza de detalhes.

O *UsefulChem*, primeira iniciativa de caderno aberto, ocorreu por meio da plataforma *blogger*, serviço oferecido pelo Google. Nessa plataforma foram depositados dados registrados em vídeos, aulas, palestras, entrevistas capítulos de livros e artigos científicos de autoria de Jean-Claude Bradley e seus colaboradores. Conforme já mencionado, o *UsefulChem* fornece dados experimentais brutos, juntamente com a interpretação do pesquisador, de forma que qualquer interessado possa facilmente reanalisar, reinterpretar e reutilizar, porém a colaboração aberta se restringia a comentários, sugestões e criticas em caixa de comentário. Sendo assim, a partir da definição de Silva (2019) para grau de estruturação de dados, os dados publicados na plataforma *Blogger* não são estruturados, pois não há um modelo de dados identificável e legível por máquina, impossibilitando que os dados sejam extraídos, transformados e processados. O caderno *UsefulChem* foi descontinuado na plataforma *re3data* e se mantém na plataforma *Blogger*, porém sem atualização.

As ferramentas tecnológicas adotadas para armazenamento e publicação dos dados foram *blogs*, *wikis* e repositórios de dados como o *re3data*, o GBIF, a *Artic Data Center* e Zenodo. A estrutura dos dados pode variar conforme a plataforma definida para a publicação, por exemplo quando se trata de postagem em *blogs ou facebook* e repositório de dados.

Os dados de pesquisa anotados nos cadernos abertos *LabScribbles* e *Openlabnotebooks* são publicados em *blogs* desenvolvidos pela plataforma *WordPress*. Os dados publicados nesses *blogs* são exportados no formato *Standard Database Format* (SDF), o qual é utilizado para arquivos de banco de dados. Os dados experimentais dos cadernos *LabScribbles* e *Openlabnotebooks* são publicados no repositório Zenodo. Nessas plataformas os dados podem ser exportados por vocabulários recomendados pelo Consorcio W3C, tais como:

- ***JavaScript Object Notation*** **(JSON)** – é um formato para troca de dados, de fácil escrita e leitura para pessoas e fácil análise e geração para máquinas. Segundo o W3C é um formato útil de serialização de dados e mensagens. É um formato legível por usuários humanos e máquina.
- ***JavaScript Object Notation for Linked Data*** **(JSON-LD)** – é um formato fundamentado em JSON para serializar dados vinculados. Destina-se principalmente a ser uma maneira de usar dados vinculados em ambientes de programação na Web,

criar serviços da Web interoperáveis e armazenar dados vinculados em mecanismos de armazenamento em JSON (W3C, 2020).

- ***Resource Description Framework / Extensible Markup Language* (*RDF/XML*) –** o RDF é um modelo padrão para intercâmbio de dados na *Web.* A XML é uma linguagem de marcação extensível que usa um formato padrão para descrição e troca de dados na *Web.* A sintaxe RDF/XML é definida para expressar um gráfico RDF como um documento XML (W3C, 2014; W3C, 2015).
- ***Data Catalog Vocabulary* (DCAT)** – é um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web. Este vocabulário permite que um editor descreva conjuntos de dados e serviços de dados em um catálogo usando um modelo e vocabulário padrão, os quais facilitam o consumo e a agregação de metadados de vários catálogos, amentando a capacidade de descoberta de conjuntos e serviços de dados (W3C, 2020).

Além destes, a plataforma Zenodo exporta dados nos formatos GeoJSON, um formato baseado no JSON, projetado para representar recursos geográficos simples; *CineStyle Color Lookup Format* (CSL) pertence à categoria de arquivos de dados criado pela Technicolor. Os arquivos com extensão CSL podem ser usados por programas distribuídos para a plataforma Windows; MARCXML é uma forma para codificação de registros MARC 21 com a linguagem de marcação XML; BibTex é uma ferramenta de formatação de bibliografia, usada em documentos LatEX. Esta ferramenta segue o mesmo conceito da distinção de conteúdo com o estilo CSS e XHTML; DataCite para fornecer identificadores persistentes aos dados de pesquisa e outros resultados de pesquisa.

Os cadernos abertos *LabScribbles* e *Openlabnotebooks*, por meio da plataforma Zenodo, adotam o *Open Graph Protocol* (OGP) para descrição dos dados. O padrão OGP foi criado para o *Facebook* e é inspirado nos padrões *Dublin Core*, *Link-rel canonical*, *Microformats* e *RDFa*. O OGP adota *tags* de metadados básicos e opcionais, conforme especificidade de cada dado ou conjunto de dados.

Os elementos básicos do OGP são: *og:type* título (*og:title*), tipo do objeto (*og:type*), URL da imagem (*og:image*) e URL do objeto que será usado como o ID identificador permanente no gráfico, por exemplo http://www.imdb.com/title/tt0117500/ (*og:url*). Alguns dos elementos opcionais são: URL para um arquivo de áudio que acompanha o objeto (*og:audio*), descrição de uma ou duas frases do objeto (*og:description*), a localização (*og:locale*), idioma (*og:locale:alternate*), endereço do site geral, quando for o caso

(*og:site_name*), e uma URL para um arquivo de vídeo que complementa as informações do objeto (*og:video*).

A figura 11, a seguir, apresenta a marcação de um registro extraído do repositório Zenodo referente a um relatório de dados sobre neurociência aberta, o qual foi descrito por metadados básicos do *Open Graph Protocol* e gerado a partir do formato RDF/XML.

Figura 11 - Registro de Dado de Pesquisa Científica em RDF/XML

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:og="http://ogp.me/ns#"
        xmlns:xhv="http://www.w3.org/1999/xhtml/vocab#">
  <rdf:Description rdf:about="https://zenodo.org/record/3350200">
    <og:title xml:lang="en">A Knowledge Mobilization Strategy for Open Neuroscience</og:title>
    <og:description xml:lang="en">Presentation by Rachel Harding, CONP Communication’s Committee Co-cha
ir, at the CONP Annual Meeting in Toronto, 2019.</og:description>
    <og:url xml:lang="en">https://zenodo.org/record/3350200</og:url>
    <og:site_name xml:lang="en">Zenodo</og:site_name>
    <xhv:license rdf:resource="http://creativecommons.org/licenses/by/4.0/legalcode"/>
  </rdf:Description>
</rdf:RDF>
```

Fonte: Zenodo (2020).

Os elementos adotados neste registro foram título (<og:title xml:lang="en">), a descrição (<og:description xml:lang="en">), a URI do site do registro (<og:url xml:lang="en">), URI do site geral (<og:site_name xml:lang="en">).

Os dados de pesquisa que os cadernos *LabScribbles* e *Openlabnotebooks* publicam seguem uma estrutura baseada em vocabulários e formatos legíveis por usuários humanos máquinas, o que facilita o uso e o compartilhamento de dados. Porém, em alguns casos, ao acessar o documento normalmente em formato de texto foram identificados gráficos em formato de foto (JPEG e PNG), o que poderia ser apresentada também em planilhas (CSV, ODS, XLS, XLSX), pois segundo Rautenberg, Souza, Dall’Agnol e Michelon (2018) os dados quando disponibilizados em tabelas, contendo linhas e colunas, facilita o seu processamento e sua recuperação por ferramentas computacionais. Além dessas fragilidades, entende-se que os objetos poderiam contemplar a descrição mais detalhada dos metadados para oferecer aos usuários melhores chances de encontrá-los. Além disso, não foi possível identificar metadados referentes à proveniência de dados, tais como fonte, declaração de proveniência e declaração de direitos os quais transmitem credibilidade ao conjunto de dados.

Silva, Santarem Segundo e Silva (2018) destacam que o novo modelo de publicação científica tem proporcionado aos dados de pesquisa maior valor enquanto ativos publicáveis

de forma autônoma e/ou ampliada. Ainda segundo os autores, para obter qualidade na recuperação dos dados de pesquisa é essencial publicar tais dados a partir de uma estrutura semântica legível por humanos e máquinas. Dessa forma, o capítulo 6 traz uma abordagem da Web Semântica a partir de suas tecnologias e conceitos do *Linked Data*, com ênfase nos vocabulários e diretrizes para publicação de dados de pesquisa cientifica.

6 WEB SEMÂNTICA E LINKED DATA

Este capítulo apresenta a Web Semântica, suas tecnologias e os conceitos do *Linked Data*, com ênfase nos vocabulários e diretrizes para publicação de dados na Web. Neste capítulo pretendeu-se identificar as tecnologias apropriadas para serem adotadas nas diretrizes semânticas para estruturação e publicação de dados anotados em cadernos de pesquisa, como meio de acelerar as descobertas e melhorar a qualidade da recuperação da informação no ambiente digital. Para tanto, buscou-se apoio teórico nos seguintes temas: 6.1 Web Semântica; 6.2 *Linked Data*; e 6.3 tecnologias para publicação de dados de pesquisa científica.

A seção 6.1 apresenta a Web Semântica a partir de seus aspectos conceituais e evolutivos, enfatizando a proposta de Tim Berners-Lee de atribuir significado às informações disponíveis na Web.

Na seção 6.2, o *Linked Data* é abordado como parte da evolução da Web Semântica e como diretrizes para implementação dessas tecnologias para publicar dados na Web, de forma a proporcionar benefícios (compreensão, interligação, descoberta, confiança, acesso, interoperabilidade, processabilidade e reuso) aos dados de pesquisa.

Na seção 6.3 são apresentadas as principais tecnologias da Web Semântica para publicação de dados de pesquisa na Web, a partir da composição de conjuntos de dados, metadados e princípios arquitetônicos da Web Semântica, tais como RDF, URI, serializações RDF, ontologias, modelo conceitual IFLA LRM. Este último pretende apoiar na definição de metadados e na descrição dos dados científicos de cadernos de pesquisa. O conjunto de tecnologias apresentado neste capítulo constituirá a base estrutural das diretrizes semânticas que esta tese propõe.

6.1 WEB SEMÂNTICA

A Web Semântica teve início em 2001, por Tim Berners-Lee com a colaboração de Hendler e Lassila, para atribuir à informação um significado bem definido, fazendo com que computadores e pessoas trabalhem em cooperação. A Web Semântica se propõe a definir uma maneira eficiente para representar dados na *World Wide Web* e proporcionar melhorias na qualidade da recuperação da informação, permitindo, conforme Santarém Segundo (2012, p. 106) "aos usuários obter resultados mais precisos e com informação mais próximas do que realmente necessitam" (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 2).

Segundo Berners-Lee, Hendler e Lassila (2001), a Web Semântica não pode ser compreendida como uma Web nova ou separada, mas uma extensão da atual. Para Laufer (2015, p. 13) "a Web evoluiu de um espaço simples para a exibição de páginas contidas em documentos estáticos, para um espaço onde diversos tipos de aplicações utilizam os navegadores como plataformas para a execução de programas". Para Berners-Lee e Hendler (2001), a Web foi projetada como um espaço de informação, com o objetivo de ser útil não apenas para a comunicação entre pessoas, mas também para que as máquinas pudessem participar e ajudar os usuários a se comunicarem. Porém, o grande obstáculo para atingir esse objetivo é o fato de a maioria das informações na Web serem projetadas exclusivamente para interpretação de usuários humanos. Sendo que, para a interpretação da máquina, os dados precisam ser estruturados e padronizados.

Portanto, a Web Semântica vem evoluindo para atribuir à informação um significado bem definido, por meio da aplicação de conceitos e tecnologias utilizados pelos computadores para desempenhar operações que darão significado às palavras no âmbito da internet. A partir desse entendimento, Berners-Lee, Hendler e Lassila (2001) destacam que na Web Semântica os computadores não apenas serão capazes de apresentar a informação contida nas páginas Web, mas também compreendê-las.

Para Berners-Lee, Hendler e Lassila (2001), a estrutura de significados para os conteúdos das páginas da Web define um ambiente onde os agentes computacionais possam processar, compartilhar, reusar e enriquecer dados, e, assim, realizar tarefas sofisticadas aos usuários.

Nesse contexto, o World Wide Web Consortium (W3C) vem desenvolvendo uma série de tecnologias e recomendações para que dados sejam publicados de maneira legível por máquinas e amplamente disponível para humanos, tornando a Web Semântica viável.

Acerca da evolução do desenvolvimento das tecnologias, tem-se que, em sua fase inicial, a Web Semântica foi construída a partir de um modelo em camadas, com algumas tecnologias que constituíram a sua base estrutural. Após a publicação do artigo *The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, em 2001, de autoria de Tim Berners-Lee, James Hendler e Ora Lassila, vários estudos e publicações em torno do tema proporcionaram uma reavaliação das camadas, surgindo novos formatos para representar os dados na Web. Nesse sentido, Santarem Segundo (2015, p. 223) reforça que, desde então, "a Web Semântica vem ganhando força e agregando novas tecnologias, funcionalidades e evoluindo para tornar real o processo de construção de ambientes semânticos". Na visão de Santarem Segundo (2014), as tecnologias como *Resource Description Framework* (RDF), *eXtensible Markup Language* (XML) e *Web Ontology Language* (OWL) e todos os conceitos que as envolvem, ganham novas versões, descritos com clareza no W3C, e tornam possível a materialização do conceito da Web Semântica.

Para Santarem Segundo (2014), as tecnologias que compõem as camadas da Web Semântica estão diretamente relacionadas ao processo de construção da informação e armazenamento das mesmas, constituindo assim ambientes que possam ter conjunto de dados ligados semanticamente.

Nas palavras de Laufer (2015, p. 70) "o horizonte que se deseja alcançar é de um banco de dados global, em que um conjunto crescente de informações possa ser acessado por um conjunto diversificado de aplicações com os mais diferentes propósitos". Nesse sentido, somente atribuir semântica às informações da Web não é o suficiente para criar um banco de dados global, faz-se necessário criar conexões entre os dados. Laufer (2015) exemplifica que milhares de pessoas publicam dados na Web em diferentes ambientes espalhados pelo mundo. Sendo assim, para a criação de um banco de dados global é necessário estabelecer o uso de uma forma padrão de conexão entre esses dados (LAUFER, 2015).

Como parte desse desenvolvimento, surge o conceito de *Linked Data* como uma maneira de publicar dados estruturados e conectados na Web, a partir de tecnologias da Web Semântica. Dessa forma, Isotani e Bittencourt (2015) complementam que a Web Semântica é a visão do Consórcio W3C sobre a Web de dados conectados, tópico relevante para se trabalhar dados de forma inteligente e automática.

6.2 LINKED DATA

Em 2006, Berners-Lee, partindo da compreensão de que "a Web Semântica não consiste apenas em colocar dados na Web. Trata-se de fazer *links*, para que uma pessoa ou

máquina possa explorar a Web dos dados" (BERNERS-LEE, 2006, tradução nossa) introduziu o conceito *Linked Data* ao propor um conjunto de princípios para criar ligações entre recursos de distintas fontes e promover a reutilização e o enriquecimento de dados na Web.

Esses princípios são:

1. Usar *Uniform Resource Identifier* - URIs para nomear itens;
2. Usar URIs HTTP para que pessoas possam procurar esses nomes;
3. Ao consultar uma URI, fornecer informações em formatos padronizados úteis [*Resource Description Framework* - RDF, SPARQL etc.];
4. Incluir sentenças com *links* para outras URIs, para permitir que itens relacionados possam ser descobertos (BERNERS-LEE, 2006).

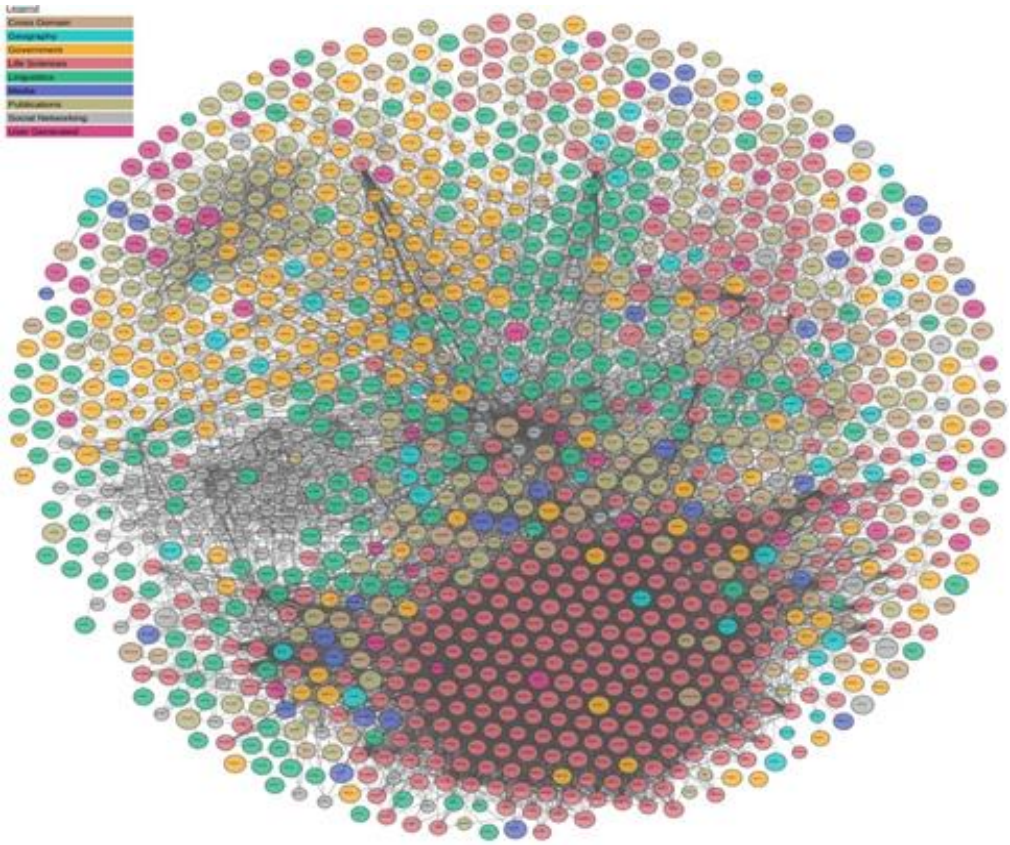
Assim, Bizer, Heath e Berners-Lee (2009, p. 2, tradução nossa) definem o *Linked Data* como "um conjunto de boas práticas para publicar e conectar dados estruturados na Web, com o intuito de criar uma Web de dados". Tais boas práticas adotam padrões reconhecidos internacionalmente e recomendados pelo W3C para construir a Web de Dados.

A adoção dessas tecnologias, seguindo as melhores práticas do *Linked Data,* alavanca a infraestrutura da Web atual com ligações e compartilhamentos de dados estruturados para o consumo humano e de máquinas, permitindo o relacionamento entre dados. Esses relacionamentos criam uma rica rede de informações interconectadas na Web, favorecendo os resultados das buscas.

Segundo Isotani e Bittencourt (2015), os dados conectados, entretanto, não necessariamente precisam ser abertos. Por exemplo, uma entidade privada pode conectar dados, mas não deixá-los abertos. Os autores salientam que, apesar de existirem iniciativas em dados conectados de forma fechada, muitas iniciativas estão se preocupando com a conexão e publicações de dados de forma aberta. De acordo com Berners-Lee (2010), o *Linked Open Data* (LOD) refere-se a dados vinculados, que são liberados sob uma licença, o que não os impedem de serem utilizados gratuitamente.

Para Santarem Segundo (2015) e Coneglian e Santarem Segundo (2017), a iniciativa mais sólida no emprego dos recursos da Web Semântica com efetividade são os *datasets* publicados por meio do LOD, o qual conta com uma grande quantidade de bases de dados estruturadas e conectadas, e vem crescendo ao longo dos anos.

Figura 12 - Nuvem de Dados Abertos Conectados (LOD)



Fonte: Diagrama de nuvem do LOD (2019).

Esta nuvem apresenta a grande expansão de dados abertos conectados e a importância dos conjuntos de dados da DBpedia (nó central de 2019, com o maior número de conexões), que contém dados representados em RDF, relativos às informações contidas nas caixas de informações dos artigos da Wikipédia. O projeto DBpedia é um importante caso de sucesso de implementação de dados conectados na Web.

A função da DBpedia é extrair conteúdos estruturados na Wikipedia e disponibilizá-los na Web. Segundo Laufer (2015), a DBpedia permite fazer consultas sofisticadas aos dados estruturados da Wikipedia e relacionar os diferentes conjuntos de dados disponíveis na Web aos artigos da Wikipedia, constituindo uma rede de ligações entre dados.

No contexto da abertura de dados, Berners-Lee (2010) sugeriu um sistema de classificação por estrelas como forma de orientação para publicar dados abertos na Web. Cada estrela representa um nível de abertura, a saber:

- **1 Estrela** - os dados são disponibilizados na Web (em qualquer formato) sob uma licença aberta.

- **2 Estrelas** - os dados são disponibilizados de forma estruturada, ou seja, legível por máquina (por exemplo, utilize o formato Excel ao invés de uma imagem escaneada).
- **3 Estrelas** - utilize formatos não-proprietários (por exemplo, utilize o CSV e não Excel).
- **4 Estrelas** - utilize URIs para identificar recursos. Isso vai colaborar para que pessoas descubram seus dados.
- **5 Estrelas** - conecte seus dados com dados de outras pessoas para prover contexto (*Linked data*).

A primeira estrela é destinada aos dados que são publicados sob licença aberta, independente do formato disponibilizado, seja em formato PDF, Word ou qualquer outro, proprietário ou não. Nesta implementação, Rautenberg, Souza, Dall1Agnol e Michelon (2018, p. 22) explicam que "os dados somente podem ser lidos, impressos, armazenados, modificados, compartilhados ou usados como dados de entrada em outros sistemas mediante o emprego de softwares para pré-processamento de dados". Assim, eles ficam restritos a um documento. Por exemplo, ao salvar dados em um arquivo PDF será necessário adotar um extrator para consumi-lo.

A segunda estrela é atribuída à publicação de dados estruturados, isto é, legível por máquinas, porém os dados ainda estão restritos em um documento proprietário. Nesse nível de abertura, os dados podem ser processados em softwares proprietários (Excel) e podem ser exportados em outros formatos, por exemplo, em XLS.

A terceira estrela é aplicada aos dados que são publicados em formato aberto não proprietário. Segundo Rautenberg, Souza, Dall1Agnol e Michelon (2018, p. 22), neste nível, 'é possível a manipulação de dados sem a necessidade de uso de um software proprietário. Um exemplo é a disponibilização na Web de um arquivo CSV em linhas separadas por vírgulas".

A quarta estrela é designada à utilização de padrões abertos recomendados pelo W3C, como RDF, SPARQL e URIs, para identificar os dados e ser identificado. Neste nível, os dados podem ser compartilhados na Web, além de permitir que outros usuários criem *links* entre os dados, facilitando o reuso de parte ou de todos eles.

A quinta estrela é atribuída aos dados que são ligados ou conectados a outros dados de fontes externas. Este nível permite a navegação e a descoberta de informações relacionadas.

Nesse caso, os dados podem ser ligados às redes semânticas e é possível enriquecer os dados gerando novos resultados e conhecimentos.

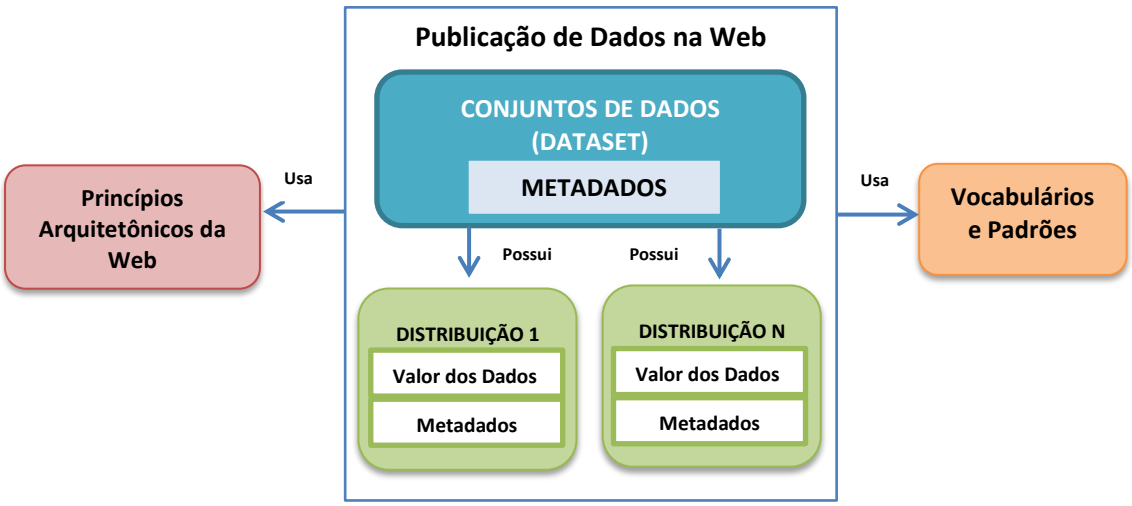
Os dados abertos conectados representam o nível ideal para publicação de dados de pesquisa na Web, pois adotam a arquitetura da Web para compartilhar dados estruturados em uma escala global, proporcionando o seu uso por diferentes pessoas e aplicações.

6.3 TECNOLOGIAS PARA PUBLICAÇÃO DE DADOS DE PESQUISA CIENTÍFICA

O Consórcio W3C recomenda um conjunto de tecnologias e melhores práticas para a publicação de dados abertos e conectados na Web. Segundo Laufer (2015, p. 33) "a ideia de agregar semântica aos dados visa facilitar o entendimento e a interoperabilidade dos dados nesse universo de informações heterogêneas, publicadas nos mais diferentes formatos e com diferentes protocolos de acesso".

Para Lóscio, Burle e Calegari (2017), as melhores práticas para dados na Web referem-se a conjuntos de dados publicados por um agente e disponibilizados para acessos e *download* em um ou mais formatos. Os dados são publicados em diferentes distribuições, as quais representam uma forma específica de disponibilizar um conjunto de dados. Essas distribuições facilitam o compartilhamento em larga escala e permite que os conjuntos de dados sejam utilizados por vários grupos de consumidores de dados, sem levar em consideração a finalidade, o público, interesse ou licença.

Figura 13 - Composição da Publicação de Dados na Web



Fonte: Lóscio, Burle e Calegari (2017, on-line).

Diante da heterogeneidade dos dados na Web, faz-se necessário fornecer informações que contribuam para a sua confiabilidade e reutilização, tais como metadados estruturais e

descritivos, acesso à informação, informações sobre qualidade e proveniência, informações sobre licença e uso. Além disso, outro aspecto a ser observado refere-se aos princípios relacionados à base arquitetônica da Web, como o uso de URIs para identificação de recursos e conexão entre dois ou mais recursos. Para promover a interoperabilidade é recomendável o uso de protocolos padronizados, vocabulários e padrões de dados (LÓSCIO; BURLE; CALEGARI, 2017).

As principais tecnologias recomendadas pelo Consórcio W3C para publicação de dados abertos e conectados na Web, de acordo com a figura 13, são descritas a seguir.

6.3.1 Resource Description Framework (RDF)

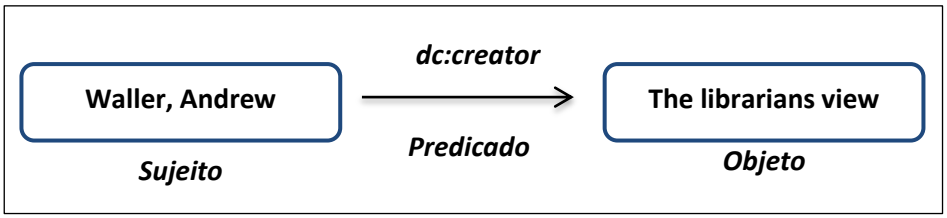
A publicação de dados abertos e conectados na Web prevê o uso de metadados para descrever, processar e localizar informações. Nesse aspecto, Laufer (2015) expõe que

> o *Resource Description Framework* (RDF) atua como uma base para o processamento dos metadados, de forma a permitir a interoperabilidade dessas descrições entre diferentes aplicações, apoiadas em padrões sobre sintaxe e a semântica dos metadados, além de um conjunto de vocabulários comuns e formas de acesso padronizadas (LAUFER, 2015, p.34).

O RDF é um modelo padrão adotado para a descrição de informações estruturadas na Web, permitindo representar a informação de um recurso de forma legível por máquinas (W3C, 2014). Para descrever a relação entre recursos e codificar o significado das etiquetas, o modelo RDF oferece uma estrutura de triplas do tipo sujeito (recurso), predicado (propriedade) e objeto (valor). O sujeito e o objeto representam dois recursos que são relacionados por um predicado. Por exemplo, Andrew Waller (sujeito) é criador (predicado) da obra *The librarians view* (objeto). Os recursos são quaisquer coisas disponíveis na Web identificadas por URIs. Assim, as três partes da tripla são vinculadas através do uso de URIs. Segundo Isotani e Bittencourt (2015), os URIs proveem uma maneira única para identificar recursos e expressar relações entre eles.

As triplas são formadas pelos elementos sujeito, predicado e objeto, como apresentado na figura 14.

Figura 14 - Representação de uma tripla RDF



Fonte: Elaborada pela autora (2020).

O sujeito “Waller, Andrew” e o objeto “The librarians view” representam dois recursos que estão relacionados pelo predicado *dc:creator*, do vocabulário Dublin Core. O conjunto de estruturas em triplas forma um grafo RDF, sendo que um grafo pode possuir múltiplas triplas em um único documento RDF. Para que o computador possa interpretar os dados representados por estas triplas é necessário que os elementos sejam descritos e referenciados por identificadores únicos de recursos. Os URIs proveem uma maneira única para identificar recursos e expressar relações entre eles. Assim, a relação RDF com URI é um conceito chave para referenciar e descrever de forma única e não ambígua os dados na Web (ISOTANI, BITTENCOURT, 2015).

Como forma de ilustrar a ideia da definição dos dados por meio de um conjunto de triplas, Laufer (2015) apresentou um modelo de dados, conhecido por modelo relacional e o seu conjunto de tabelas relacionadas. Nesse modelo, o autor apresenta uma tabela com informações sobre um livro, por exemplo, ISBN, título, autor e editora. Cada linha da tabela é composta por informações referentes a um determinado livro. Cada um desses livros é um recurso. Cada coluna da tabela define um tipo de propriedade ou predicado relacionado aos livros. Cada célula da tabela define uma tripla.

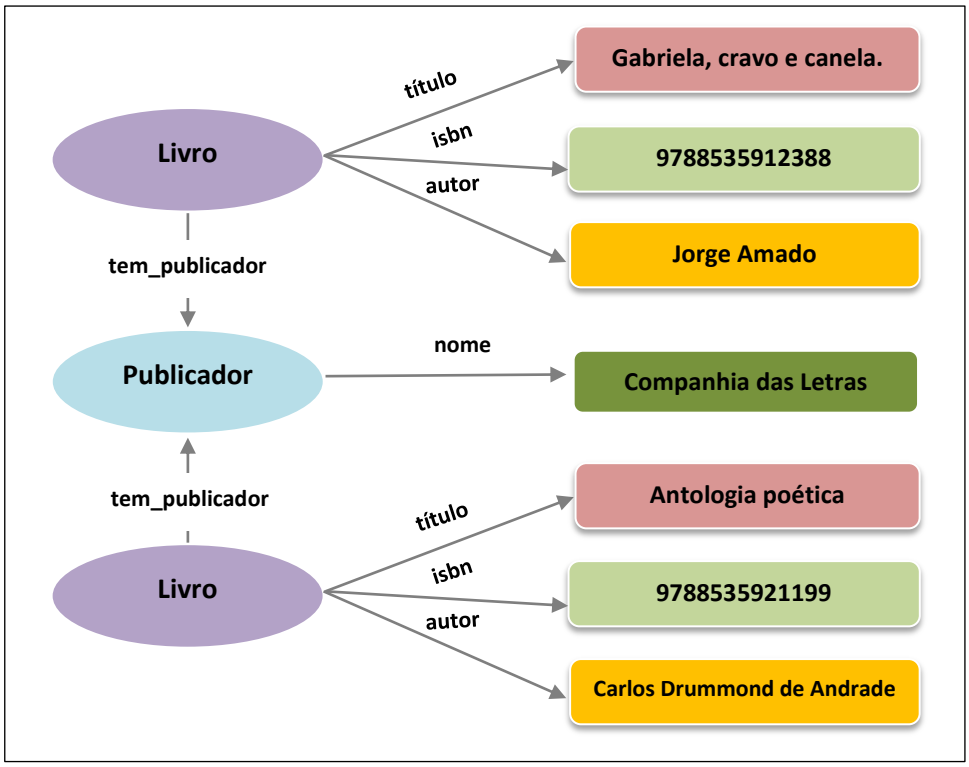
Tabela 01 - Tabela Relacionada

ISBN	Título	Autor	Editora
9788535912388	Gabriela, Cravo e Canela	Jorge Amado	Companhia das Letras
9788501067340	Vidas Secas	Graciliano Ramos	Companhia das Letras
9788535921199	Antologia Poética	Carlos Drummond de Andrade	Companhia das Letras

Fonte: Adaptada de Laufer (2015).

A partir da tabela relacionada, uma tripla pode ser representada como uma espécie de grafo dirigido do sujeito para o objeto, como o exemplo da figura 15. O grafo RDF pode compor triplas de três propriedades (título, isbn e autor) de um mesmo recurso (livro), além de considerar a relação do livro com o seu publicador. O grafo pode ampliar se considerar dois livros do mesmo publicador e cada qual com suas propriedades, conforme a figura 15.

Figura 15 - Grafo RDF de vários recursos



Fonte: Laufer (2015, p.38).

Cada um desses recursos e propriedades podem ser identificados por URIs como forma de garantir a exatidão das informações representadas e conectadas para oferecer aos usuários a informação, da qual realmente necessita.

6.3.2 *Uniform Resource Identifier* (URIs)

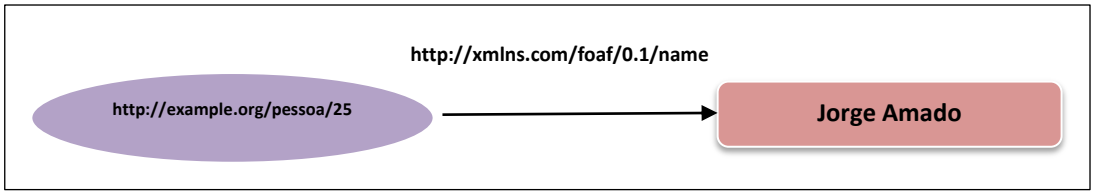
O URI é um padrão responsável por identificar um recurso físico ou abstrato de maneira única e global (SANTAREM SEGUNDO, 2010). No cenário da Web todos os recursos e propriedades devem ser identificados por um URI, de modo a obter uma semântica única e consistência dos dados.

Nesse contexto, Serra (2019) observa que uma das complexidades da organização da informação no ambiente Web refere-se à ambiguidade e à falta de consistência dos dados, tornando-se necessário que coisas, pessoas e conceitos, enquanto recursos da Web, tenham um identificador único para que os dados possam prover um resultado desejado. Outro aspecto relacionado com a falta de consistência dos dados é a nomenclatura empregada nas propriedades. Um exemplo que pode ser utilizado para ilustrar essa situação é a definição do termo título adotado na tabela relacional. Em uma determinada organização adota-se ‘title’,

em outra organização pode-se utilizar ‘nome’, ‘nome da obra’ ou ainda ‘ttl’. Para Laufer (2015) a forma particular (cada organização adota a sua nomenclatura) de atribuir nomenclaturas pode ocorrer em perdas semânticas.

Em RDF a forma de identificar recursos e propriedades, de forma única e universal, é utilizar URIs, mas especificamente com HTTP URIs (LAUFER, 2015, p. 39). Segundo Berners-Lee (2002, on-line), URIs HTTP definem uma rede de objetos de informação, os quais normalmente [...] “são coisas que leve alguma mensagem e pode ser representado para uma maior ou menor autenticidade, em *bits”*. Para o autor, usa-se URIs HTTP para que pessoas possam encontrar nomes na Web. Dessa forma, a figura 16 ilustra esse uso.

Figura 16 - Uso de URIs. Exemplo de uso do foaf:name



Fonte: Laufer (2015, p. 38).

Na ilustração 16 adotou-se o vocabulário FOAF para descrever o autor. Segundo Lóscio, Burle e Calegari (2017, on-line), o uso de vocabulários de referência para a definição de propriedades de domínios específicos é uma maneira de garantir a semântica, um exemplo de vocabulário é o FOAF que descreve pessoas, suas atividades e suas relações com outras pessoas e objetos.

As declarações em RDF podem ser apresentadas em formato de serialização, conforme é tratado na seção 6.3.3.

6.3.3 Serializações RDF

O modelo RDF oferece várias notações, entre elas incluem *Resource Description Framework / Extensible Markup Language* (RDF/XML), *Terse RDF Triple Language* (Turtle), N-Triples e *JavaScript Object Notation for Linked Data* **(**JSON-LD).

A notação RDF/XML foi a primeira serialização feita para RDF. Laufer (2015, p. 41) declara que essa notação, inicialmente, tinha a vantagem pelo fato das linguagens de programação ter mais suporte para XML, apesar de ser considerada bastante difícil para leitura de pessoas. De acordo com o autor, a notação RDF/XML permite fazer uso de

namespaces XML para evitar o uso de URIs completas, o que torna os URIs menos extensos. Um exemplo da notação RDF/XML foi apresentado na figura 11.

A sintaxe Turtle permite que um grafo RDF seja completamente escrito em um formato de texto compacto, com abreviações para padrões de uso e tipos de dados comuns (W3C, 2014). Por ser apresentada em formato de texto, a notação Turtle é considerada simples para leitura de pessoas. Como forma de ilustrar, apresenta-se na figura 17, a sintaxe Turtle do conjunto de dados de pesquisa apresentado no exemplo da figura 11 (Registro de Dado Científico em RDF/XML), extraído da plataforma Zenodo.

Figura 17 - Registro de Dados de Pesquisa Científica em Turtle

```
@prefix og: <http://ogp.me/ns#> .
@prefix xhv: <http://www.w3.org/1999/xhtml/vocab#> .
<https://zenodo.org/record/3350200>
  og:title "A Knowledge Mobilization Strategy for Open Neuroscience"@en ;
  og:description "Presentation by Rachel Harding, CONP Communication’s Committee Co-chair, at the CONP A
nnual Meeting in Toronto, 2019."@en ;
  og:url "https://zenodo.org/record/3350200"@en ;
  og:site_name "Zenodo"@en ;
  xhv:license <http://creativecommons.org/licenses/by/4.0/legalcode> .
```

Fonte: Zenodo (2020), gerado a partir da ferramenta EasyRdf.

A notação Turtle fornece níveis de compatibilidade com o formato N-Triples, bem como a sintaxe de padrão triplo de recomendação SPARQL (W3C, 2014).

A notação N-Triples é um subconjunto de Turtle e usa um formato baseado em linhas para codificar um grafo RDF, sendo utilizada como um formato de troca de dados RDF de maior legibilidade por máquinas. Rautenberg, Souza, Dall’Agnol e Michelon (2018) explicam que na notação N-Triples, os URIs completos são apresentados dentro dos delimitadores de maior (>) e menor (<) e *strings* entre aspas duplas (“ ”). Cada tripla ocupa uma linha com um ponto ao final.

Figura 18 - Registro de Dados de Pesquisa Científica em N-Triples

```
<https://zenodo.org/record/3350200> <http://ogp.me/ns#title> "A Knowledge Mobilization Strategy for Open N
euroscience"@en .
<https://zenodo.org/record/3350200> <http://ogp.me/ns#description> "Presentation by Rachel Harding, CONP C
ommunication\u00E2\u0080\u0099s Committee Co-chair, at the CONP Annual Meeting in Toronto, 2019."@en
.
<https://zenodo.org/record/3350200> <http://ogp.me/ns#url> "https://zenodo.org/record/3350200"@en .
```

<https://zenodo.org/record/3350200> <http://ogp.me/ns#site_name> "Zenodo"@en .
<https://zenodo.org/record/3350200> <http://www.w3.org/1999/xhtml/vocab#license> <http://creativecommons.org/licenses/by/4.0/legalcode> .

Fonte: Zenodo (2020), gerado a partir da ferramenta EasyRdf.

A serialização JSON-LD é um formato baseado em JSON para *Linked Data*. Para Laufer (2015, p. 43) um elemento central de JSON-LD é a ideia de contexto.

> Quando duas pessoas se comunicam em um ambiente compartilhado, existe um contexto de conhecimento mútuo que permite que os indivíduos usem termos abreviados, um vocabulário próprio, para se comunicar mais rapidamente, mas sem perder a precisão. Um contexto em JSON-LD funciona da mesma forma. Ele permite que duas aplicações usem termos abreviados, termos particulares, para se comunicar com mais eficiência, mas sem perder a precisão.

Para Kellogg, Champin e Longley (2020) a serialização JSON-LD destina-se principalmente a ser uma maneira de usar dados vinculados em ambientes de programação baseados na Web e criar serviços interoperáveis. Isotani e Bittencourt (2015) ressaltam que este formato é intuitivo para programadores já familiarizados com a sintaxe JSON.

Figura 19 - Registro de Dados de Pesquisa Científica em JSON-LD

[{"@id":"http://creativecommons.org/licenses/by/4.0/legalcode"},{"@id":"https://zenodo.org/record/3350200","http://ogp.me/ns#title":[{"@value":"A Knowledge Mobilization Strategy for Open Neuroscience","@language":"en"}],"http://ogp.me/ns#description":[{"@value":"Presentation by Rachel Harding, CONP Communicationâ€™s Committee Co-chair, at the CONP Annual Meeting in Toronto, 2019.","@language":"en"}],"http://ogp.me/ns#url":[{"@value":"https://zenodo.org/record/3350200","@language":"en"}],"http://ogp.me/ns#site_name":[{"@value":"Zenodo","@language":"en"}],"http://www.w3.org/1999/xhtml/vocab#license":[{"@id":"http://creativecommons.org/licenses/by/4.0/legalcode"}]}]

Fonte: Zenodo (2020), gerado a partir da ferramenta EasyRdf.

Além da descrição dos recursos é necessário estabelecer relações entre eles. Para isso, a Web Semântica faz uso de ontologias, as quais estão abordadas na seção 6.3.4.

6.3.4 Ontologias

Para Rautenberg, Souza, Dall'Agnol e Michelon (2018, p. 29) as ontologias estabelecem relações e possibilitam a descrição de recursos na forma de classes, propriedades e instâncias. Para os autores, "na Web Semântica, as ontologias contribuem para o compartilhamento de dados, por definirem os vocabulários comuns para representação de dados conectados".

Segundo Patrício (2012), as ontologias são compostas por uma taxonomia, que define as classes de objetos e as relações entre eles, e por regras de inferência as quais permitem que os programas computacionais manipulem os termos de forma mais eficiente do que os seres humanos. Duas linguagens adotadas na Web Semântica para o desenvolvimento de ontologias são RDF *Schema* (RDFS) e *Web Ontology Language* (OWL).

De acordo com Isotani e Bittencourt (2015, p.69) "o RDFS é um vocabulário para modelagem de dados que amplia a expressividade do RDF para promover mecanismos de descrição de taxonomias e suas propriedades". Esse vocabulário fornece um sistema básico de classes e propriedades e indica como são usados em conjunto, sendo que o conceito classe é utilizado para descrever recursos em um documento RDF e a propriedade define relações entre sujeitos e objetos.

A OWL é uma ontologia mais extensa e expressiva do que a RDFS, usada para descrever e definir termos dentro de um determinado domínio de interesse ou assunto e para descrever e definir as relações entre eles. Na OWL existem três tipos de entidades: classes, relacionamentos ou propriedades e instâncias. As instâncias representam os recursos (também chamados de indivíduos); as classes definem conjuntos de instâncias, de indivíduos; e as propriedades representam uma associação entre indivíduos das classes.

Outras ontologias desenvolvidas e consolidadas são o *Friend of a Friend* (FOAF) e o *Simple Knowledge Organization System* (SKOS). O FOAF tem por objetivo descrever relacionamentos entre pessoas e informações na Web. O SKOS desenvolve especificações e padrões para utilização de organização do conhecimento, como tesauros, esquemas de classificação, taxonomias e cabeçalhos de assuntos no âmbito da Web Semântica, com capacidade de expressar suas relações hierárquicas e associativas (RYAN; GRANT; COLLINS; STEFAN; LOPES, 2017).

Para consultas e acesso a conjuntos de dados descritos em RDF, destaca-se a interface SPARQL. De acordo com a W3C (2008), o SPARQL refere-se a um conjunto de especificações que fornece linguagens e protocolos para recuperar e manipular dados armazenados em RDF.

Além desses padrões, Riva, Le Boeuf e Žumer (2017) destacam o modelo conceitual IFLA *Library Reference Model* (LRM) que faz uso dos princípios de *Linked Data* para colaborar na definição de metadados e na descrição dos dados na Web, com o foco no usuário.

6.3.5 Modelo Conceitual IFLA LRM

Na visão de Mey e Silveira (2009, p. 8) "[a representação da informação] se fundamenta nos relacionamentos entre os registros do conhecimento, estabelecidos de forma a criar alternativas de escolha para os usuários". As autoras trazem reflexões sobre a importância de relacionamentos entre registros no contexto da descrição e recuperação da informação no ambiente Web para servir de conveniência ao usuário com acesso a informação.

Nesse contexto, novos modelos conceituais bibliográficos estão sendo desenvolvidos para oferecer consistência na descrição de coleções de dados e estabelecer relacionamentos entre eles. Entre os modelos destacam-se o estudo o IFLA *Library Reference Model* (IFLA LRM) que originou da harmonização dos principais modelos conceituais do tipo entidade-relacionamento representados pela família FR, composta pelos modelos conceituais *Functional Requirements of Bibliographic Records* (FRBR), do português Requisitos Funcionais para Registros Bibliográfico, *Functional Requirements for Authority Data* (FRAD), do português Requisitos Funcionais para Dados de Autoridades e *Functional for Subject Authority Data* (FRSAD), do português Requisitos Funcionais para Dados de Assunto.

O modelo conceitual IFLA LRM foi intitulado inicialmente como FRBR LRM até 2016, quando teve a aprovação do relatório final do Modelo de Referência para Dados Bibliográficos pela IFLA, em 2017. O IFLA LRM pretende ser um modelo conceitual único, apoiado em uma estrutura entidade-relacionamento de alto nível, focado nas tarefas do usuário e característica hierárquica. Esse modelo teve algumas entidades herdadas ou derivadas dos modelos anteriores (RIVA; LE BOEUF; ŽUMER, 2017). As entidades do modelo FRBR são essenciais e consideradas a base do modelo IFLA LRM.

Para Ryan (2015), o FRBR é um modelo de relacionamento de entidade que descreve o universo bibliográfico em relação às necessidades dos usuários e os recursos disponíveis. Segundo Welsh e Batley (2012), em FRBR, os atributos e relacionamentos dentro de um catálogo são mapeados para as tarefas do usuário, os quais formam os objetivos centrais de um registro. As tarefas dos usuários são:

- **Encontrar** – um recurso em uma coleção utilizando seus atributos e suas relações.
- **Identificar** – um recurso ou confirmar se esse recurso corresponde ao procurado ou ainda, distinguir um recurso entre outros semelhantes.

- **Selecionar** – um recurso que seja apropriado às necessidades do usuário no que se refere ao conteúdo, suporte etc.
- **Adquirir ou obter** – acesso a um recurso, ou seja, fornecer informação que permite adquirir um recurso por compra, empréstimo etc.
- **Navegar** – por um catálogo, pela ordenação lógica da informação bibliográfica e apresentação de vias claras para transitar, incluindo a apresentação das relações entre obras, expressões, manifestações e exemplares (WELSH; BATLEY, 2012).

Fusco (2011, p. 20) explica que o modelo E-R, proposto por Chen (1990), "define uma representação de informações para modelagem de banco de dados baseada em entidades, atributos e relacionamentos entre as entidades".

> O Modelo E-R é uma metodologia de construção de modelos conceituais que se baseia na percepção do domínio do cenário como um conjunto de objetos básicos, chamados entidades e o relacionamento entre eles. As entidades são descritas por meio de seus atributos. O número das entidades às quais uma outra entidade se relaciona é determinado pelo mapeamento das cardinalidades (SILBERSCHATZ; *et al*., 1999 *apud* FUSCO, 2011, p. 100).

"O modelo entidade-relacionamento é uma metodologia que mapeia as necessidades do sistema, por isso é importante que se conheçam o domínio e as necessidades do usuário para que se construa um modelo conceitual com requisitos sólidos" (COAD; YOURDON, 1991 *apud* ALVES, 2010, p. 81).

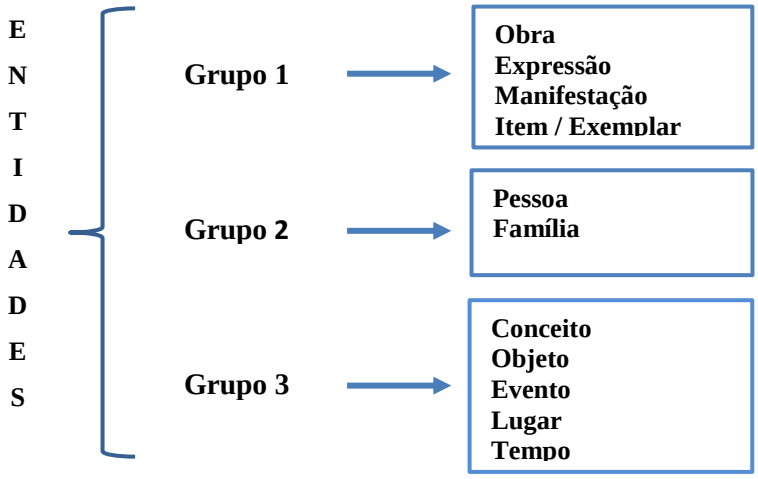
A proposta da família FRBR foi fornecer um quadro estruturado, claramente definido, para relacionar dados registrados em registros bibliográficos às necessidades dos usuários desses registros, e recomendar um nível básico de funcionalidade para registros criados por entidades bibliográficas nacionais (IFLA, 1998, p. 7).

> A utilização dos conceitos estabelecidos pelos FRBR proporcionará o estabelecimento da recuperação da informação de forma integrada, ou seja, tornará possível a recuperação de uma obra (obra - substitui a palavra livro no conceito do conteúdo) em todas as suas expressões e todos os itens (termo item é utilizado para identificar o objeto físico ou não) em que tiver sido manifestada (manifestação é a forma de apresentação de uma obra) (IFLA, 1998, p. 7).

São identificadas dez entidades nos FRBR, que, por sua vez, dividem-se em três grupos, conforme o que é apresentado na figura 20 - Entidades dos FRBR. As entidades encontradas no grupo 1 são aquelas que representam os produtos do trabalho intelectual ou artístico; no grupo 2 estão aquelas que representam os responsáveis pelo conteúdo, produção, disseminação e, ou, guarda das entidades do grupo 1; e, no grupo 3 estão as entidades que

representam os assuntos de uma obra (porém, os grupos 1 e 2 também possam representar um assunto de uma obra) (MEY; SILVEIRA, 2009).

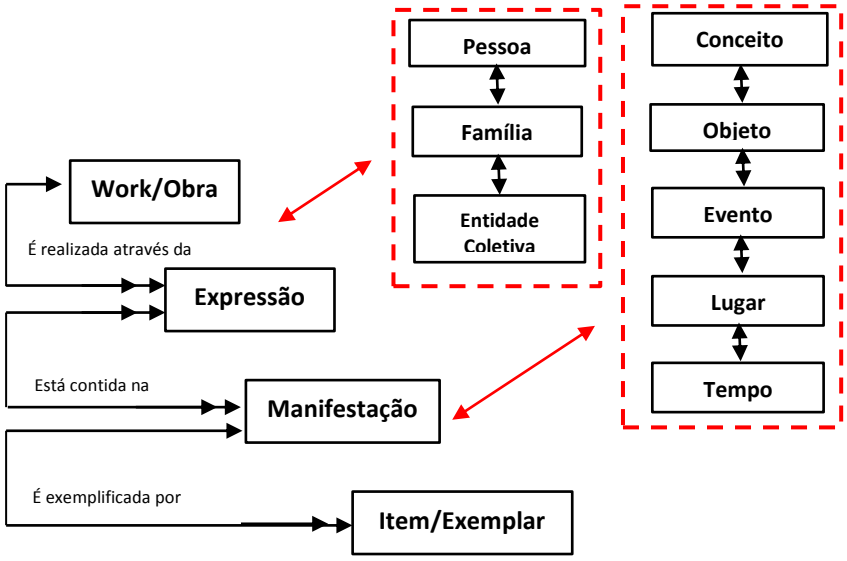
Figura 20 - Entidades dos FRBR



Fonte: Adaptado Mey e Silveira (2009, p. 19).

A figura 21 explicita as relações entre os conceitos pertencentes aos grupos 1, 2 e 3.

Figura 21 – Relações entre as entidades



Fonte: Adaptado a partir da IFLA (2016).

As entidades do grupo 1 – obra, expressão, manifestação e item/exemplar – representam os produtos do trabalho intelectual ou artístico. Segundo Mey e Silveira (2009, p. 19), os conceitos do grupo 1 podem ser definidos do seguinte modo:

- **Obra**: criação intelectual ou artística distinta (é o conteúdo intelectual em si,

independentemente de seu suporte ou de sua forma). Uma obra também pode ser o assunto de uma obra.

- **Expressão**: realização intelectual ou artística de uma obra (é a forma como se expressa o conteúdo intelectual - compreende traduções, interpretações de uma obra musical determinada, entre outros). Uma expressão também pode ser o assunto de uma obra.
- **Manifestação**: materialização de uma expressão de uma obra (é a representação de todos os objetos físicos que possuem as mesmas características, tanto de conteúdo intelectual como de forma física). Abrange um amplo leque de materiais (manuscritos, livros, periódicos, mapas, cartazes, registros sonoros, filmes, vídeos etc.), e pode ser considerado o suporte físico de uma expressão ou obra. A manifestação pode ser o assunto de uma obra.
- **Item**: exemplificação única de uma manifestação (é o objeto físico que permite que usuário acesse o conteúdo intelectual ou artístico de uma expressão e de uma obra). Compreende o objeto existente em um lugar determinado (mesmo o ciberespaço) e pode constituir-se de vários volumes. Um item também pode ser o assunto de uma obra. Segundo a IFLA (2016, p. 7, tradução nossa), "a tradução de Item por Exemplar foi feita respeitando o vocabulário usado na tradução dos Princípios de 2009".

As entidades do grupo 2 – pessoa, família e entidade coletiva – representam os responsáveis pelo conteúdo, produção, disseminação e, ou, guarda das entidades do primeiro grupo. Segundo Mey e Silveira (2009, p.20) e IFLA (2009), os conceitos do grupo 2 podem ser definidos da seguinte maneira:

- **Pessoa**: um indivíduo relacionado à criação ou realização de uma obra ou de uma expressão, ou assunto de uma obra.
- **Família**: duas ou mais pessoas relacionadas pelo nascimento, casamento, adoção ou estado legal semelhante, ou que de outro modo se apresentam como uma família.
- **Entidade coletiva**: uma organização ou grupo de indivíduos, de caráter permanente ou temporário, ou um governo territorial, que age unificadamente e se identifica por um nome.

As entidades do grupo 3 – conceito, objeto, evento, lugar e tempo – representam os assuntos de uma obra. Mey e Silveira (2009, p. 22), definem os conceitos do grupo 3:

- **Conceito**: uma noção ou ideia abstrata, sempre assunto de uma obra.
- **Objeto**: uma coisa material, móvel ou imóvel, sempre assunto de uma obra.
- **Evento**: uma ação ou ocorrência, como eventos históricos, épocas e períodos de tempo, sempre assunto de uma obra.
- **Lugar**: um local, enquanto assunto de uma obra.
- **Tempo**: datas, enquanto assunto de uma obra.

De acordo com Oliver (2011, p. 26) "os relacionamentos do modelo FRBR são fundamentais na identificação e mapeamento das relações entre as entidades. As relações desempenham um papel importante no auxílio às tarefas do usuário". Oliver (2011) complementa que esse relacionamento é indispensável à navegação no universo bibliográfico.

Além do modelo conceitual FRBR para dados bibliográficos, a família de modelos conceituais FR incluiu os Requisitos Funcionais para Dados de Autoridade (FRAD) e os Requisitos Funcionais para Dados de Autoridade do Assunto (FRSAD).

O FRAD é uma extensão do modelo FRBR, cobrindo dados de autoridades, os quais são elaborados para controlar as formas autorizadas e variantes do nome e identificadores utilizados como pontos de acesso. O FRAD foi construído em 2009 a partir da mesma técnica de análise de entidade utilizada no FRBR, com foco nas entidades do grupo 2 - pessoa, entidade coletiva, e família, como descrito acima, nesta mesma seção.

As tarefas desempenhadas pelos usuários são:

- **Encontrar** uma entidade ou conjunto de entidades correspondentes a um critério determinado, ou explorar o universo de entidades bibliográficas utilizando seus atributos e relações;
- **Identificar** uma entidade ou validar a forma do nome a ser usado como ponto de acesso controlado;
- **Contextualizar ou** localizar uma pessoa, entidade coletiva, obra etc. no contexto; esclarecer o relacionamento entre duas ou mais pessoas, entidades coletivas, obras etc.; ou esclarecer o relacionamento entre uma pessoa, entidade coletiva etc. e o nome pelo qual essa pessoa, entidade coletiva etc. é conhecida (por exemplo, o nome utilizado na religião versus o nome secular)
- **Justificar ou** documentar a razão pela qual o criador dos dados de autoridade escolheu o nome ou a forma do nome na qual o ponto de acesso controlado está baseado.

O FRSAD se dedica a questões relacionadas aos dados de autoridade de assunto e a

investigar o uso direto e indireto de autoridades de assunto por uma ampla gama de usuários. (IFLA, 2010, p. 9). As tarefas do usuário desse modelo consistem em:

- **encontrar** um ou mais assuntos e/ou suas denominações que correspondam ao critério de busca estipulado, utilizando atributos e relacionamentos;
- **identificar** um assunto e/ou sua denominação baseado em seus atributos ou relacionamentos (ou seja, distinguir entre dois ou mais assuntos ou denominações com características similares e confirmar que o assunto ou a denominação apropriada foi encontrada);
- **selecionar** um assunto e/ou uma denominação apropriada às necessidades do usuário (ou seja, escolher ou rejeitar com base nos requisitos e nas necessidades do usuário);
- **explorar** os relacionamentos entre assuntos e/ou suas denominações (ou seja, explorar os relacionamentos para entender a estrutura de um domínio de assunto e sua terminologia).

É importante destacar que o FRSAD define as entidades **thema** e **nomen**, que se relacionam à entidade obra, definida no FRBR: o *thema* é definido como qualquer entidade usada como assunto de uma obra; e *nomen* é definido como qualquer sinal ou sequência de sinais (caracteres alfanuméricos, símbolos, som etc.) pelos quais um tema é conhecido, referido ou tratado. Essas entidades serão mencionadas na discussão sobre definição e redifinição do modelo conceitual IFLA LRM, no decorrer deste texto.

Os três modelos conceituais da família FR (FRBR, FRAD e FRSAD) foram criados em uma estrutura de modelagem de relacionamento entre entidades. Contudo, o que se observou foi que, na prática, adotam diferentes pontos de vista e diferentes soluções para os mesmos problemas. Então, o grupo de trabalho da IFLA decidiu por combinar e consolidar a família FR em um único modelo coerente para esclarecer o entendimento de modelo geral e remover barreiras à sua adoção (RIVA; LE BOEUF; ŽUMER, 2017, p. 5).

A base de implementação desse modelo geral é de alto nível e, como tal, para qualquer aplicação prática precisará determinar um nível apropriado de precisão, sendo necessário expandir o contexto do modelo ou omitir alguns de seus elementos. No entanto, como já mencionado, o LRM requer que a estrutura básica das entidades e dos relacionamentos se mantenha para garantir a fidelidade ao modelo.

Sendo assim, define que a base do modelo seja as entidades do grupo 1 do FRBR e seus relacionamentos. A IFLA declara que as outras entidades e relacionamentos não são

obrigatórias e fica a caráter da necessidade de cada aplicação, bem como realizar adaptações, caso sejam necessárias, para cada especificidade. Ainda assim, são consideradas implementações do IFLA LRM. Para Riva, Le Boeuf e Žumer (2017), o IFLA LRM é um modelo lógico com foco na organização estrutural de dados e, por isso, está mais longe da prática que os outros modelos. Ao mesmo tempo, ele se adianta ao adotar a forma necessária para sua utilização em aplicações *Linked Data*.

Esse modelo tem como foco as tarefas do usuário, as quais possuem as seguintes definições, segundo Riva, Le Boeuf, Žumer (2017):

- **Encontrar** refere-se a trazer informações sobre um ou mais recursos de interesse pela pesquisa sob qualquer critério relevante.
- **Identificar** compreende a maneira de explicitar o entendimento da natureza dos recursos encontrados para distinguir entre os recursos similares.
- **Selecionar** trata da capacidade de determinar a adequação dos recursos encontrados e de aceitar ou rejeitar recursos específicos.
- **Obter** refere-se à capacidade de acessar o conteúdo do recurso
- **Explorar** remete à descoberta de recursos usando as relações entre eles e também colocar tais recursos em um contexto.

As referidas tarefas são possíveis mediante a composição da estrutrura do modelo. A estrutura do IFLA LRM é composta por entidades, atributos e relacionamentos, sendo que as entidades são categorias abstratas de objetos conceituais conectados por relacionamentos e suas características descritas por atributos. Essas entidades definem a estrutura do modelo e funcionam como nós, enquanto os relacionamentos conectam entidades entre si (RIVA; LE BOEUF; ŽUMER, 2017), conforme hierarquia das entidades apresentadas no quadro 14.

Quadro 14 - Hierarquia das Entidades do modelo IFLA LRM

Primeiro Nível	**Segundo Nível**	**Terceiro Nível**
LRM-E1 Res		
--	LRM-E2 Obra	
--	LRM-E3 Expressão	
--	LRM-E4 Manifestação	
--	LRM-E5 Item	
--	LRM-E6 Agente	
--	--	LRM-E7 Pessoa
--	--	LRM-E8 Agente coletivo
--	LRM-E9 Nomen	
--	LRM-E10 Lugar	
--	LRM-E11 Intervalo de tempo	

Fonte: Riva, Le Boeuf, Žumer (2017, p. 19, tradução nossa).

Cada elemento do modelo recebe o prefixo “LRM-”, uma letra correspondente ao tipo de elemento (E = entidade; A = atributo; R = relacionamento) e um número sequencial. Para os atributos, o número da entidade para a qual o atributo está definido é inserido antes da letra "A" (que significa atributo), em número sequencial, sendo que a numeração sequencial é reiniciada para cada entidade.

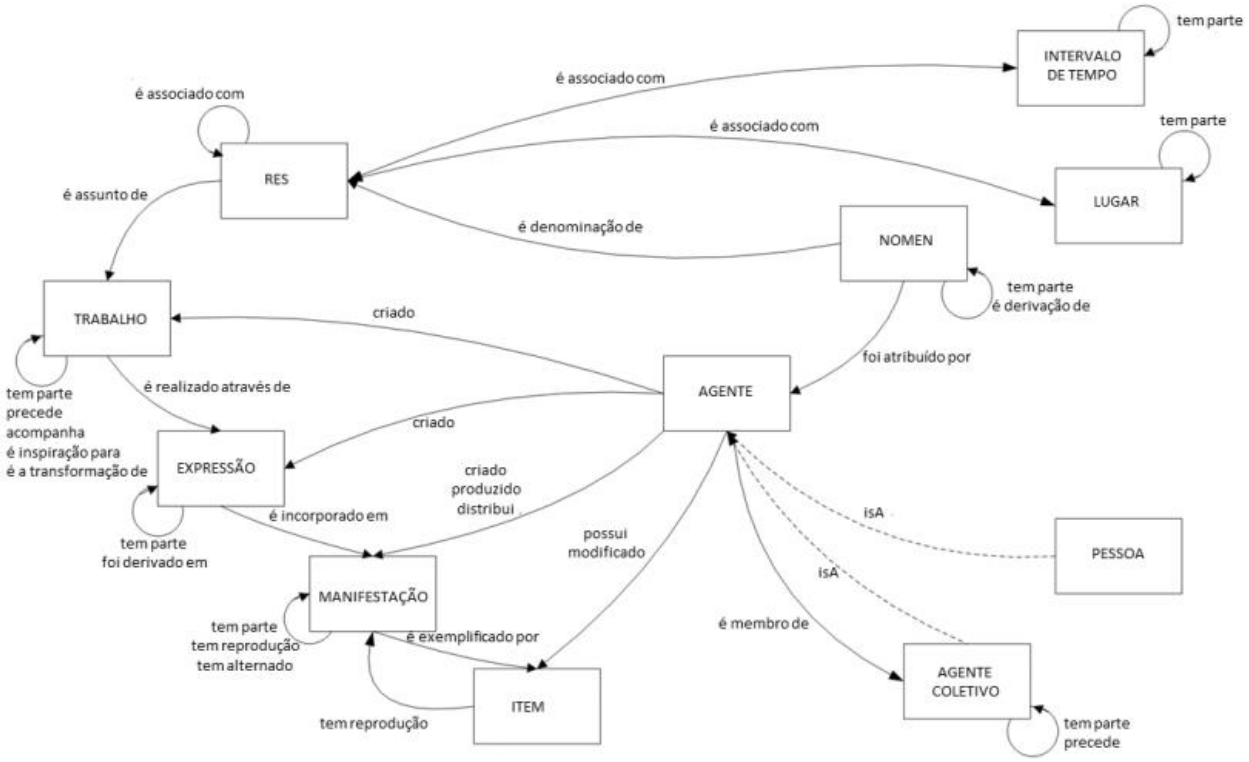
A LRM –E1 Res é a única entidade de primeiro nível Res (coisa), a qual é superior a todas as outras entidades. Uma entidade pode ser declarada como uma superclasse de todas as outras e, em seguida, têm relação de subclasse. Qualquer instância de uma entidade de subclasse também é uma instância da superclasse. Isso faz parte da estrutura de melhorias dos modelos de relacionamento entre entidades e pode ser expresso como **isA** (é um), por exemplo, a pessoa da entidade é subclasse do Agente da entidade, o que pode ser expresso como *person isA agent* (pessoa é um agente).

A entidade LRM-E2 corresponde à obra do recurso, ou seja, ao conteúdo artístico ou intelectual de uma criação distinta. A LRM-E3 refere-se à expressão que é uma combinação distinta de sinais os quais transmitem o conteúdo artístico ou intelectual da obra. A LRM-E4 corresponde à manifestação, um conjunto de todos os suportes para os quais se assume que compartilhem as mesmas características relacionadas a conteúdos artísticos ou intelectuais de forma física. Esse conjunto é definido tanto pelo conteúdo global como pelo plano de produção dos seus suportes. A entidade LRM-E5 corresponde ao item, um objeto ou objetos que contém símbolos com o objetivo de transmitir um conteúdo intelectual ou artístico. A entidade LRM-E6 corresponde ao agente, uma entidade capaz de ações intencionais, de receber direitos e de ser responsabilizada por suas ações. Essa entidade se desmembra nas entidades de terceiro nível LRM-E7, para indicar pessoa e LRM-E8 para indicar o agente coletivo ou entidade coletiva, como é conhecida no contexto catalogação.

As três últimas entidades apresentadas no quadro 14 referem-se às entidades de segundo nível. Portanto, tem-se a LRM-E9 para *Nomen*, uma associação entre uma entidade e uma denominação que se refere a ela. A entidade LRM-E10 para lugar, uma área delimitada no espaço. E, por último, a entidade LRM-E11 para indicação de intervalo de tempo, uma dimensão temporal com inicio, fim e duração.

Na figura 22 é possível visualizar os relacionamentos entre as entidades do modelo IFLA LRM.

Figura 22 - Visão geral dos relacionamentos do modelo IFLA LRM



Fonte: Riva, Le Boeuf, Žumer (2017, p. 86, tradução de Silva, 2017, p. 106).

O modelo IFLA LRM declara as relações de maneira abstrata para facilitar a implementação de projetos específicos. O modelo apresenta a relação do tipo associação, por meio da entidade Res, a qual é o nível superior e geral do modelo, válida para todas as entidades do universo bibliográfico. Em termos gerais, se definem especificações para que a semântica seja mais precisa, ou seja, outros relacionamentos adicionais necessários para uma implementação específica podem ser definidos como refinamentos adicionais do relacionamento superior (RIVA; LE BOEUF; ŽUMER, 2017).

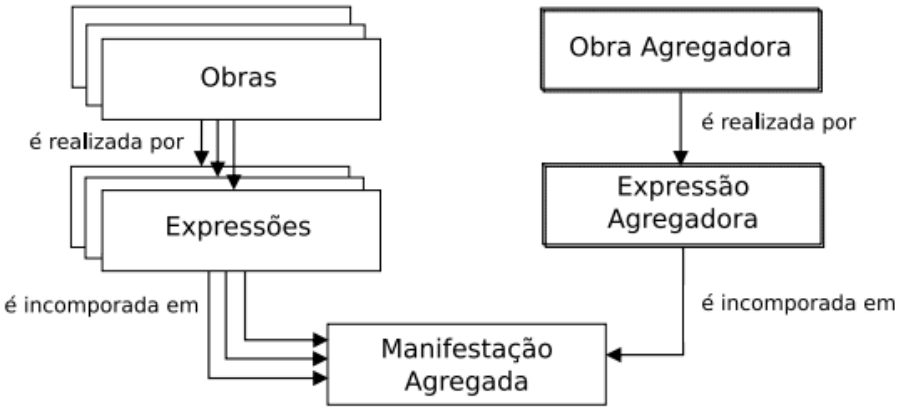
As relações do grupo 1 do FRBR (obra, expressão, manifestação e item) continuam as mesmas e obrigatórias na implementação do modelo, ainda assim, a definição de novas entidades é encorajada pelo modelo. Padron (2019) analisa a definição e redefinição de entidades dos modelos FRAD e FRSAD pelo IFLA LRM ao explicitar que

> a entidade Res redefine a entidade Thema do modelo FRAD. Foi adicionada a entidade Agente como superclasse das entidades Pessoa e Agente Coletivo, que, por sua vez, representa as entidades Entidade Coletiva e a Família do FRBR. A entidade Nomen é uma fusão das entidades Nomen do FRSAD, Nome do FRAD e de ponto de acesso controlado. Foi adicionada a entidade Intervalo de Tempo, inexistentes nos modelos anteriores (PADRON, 2019, p. 109 *apud* BIANCHINI, 2017).

Os autores destacam a adição da entidade agente, que declara um relacionamento de pertencimento entre as entidades agente coletivo e pessoa. Nesse sentido, é adicionada uma indicação da hierarquia *isA* entre a entidade Agente e suas subclasses (pessoa e agente coletivo) (RIVA; LE BOEUF; ŽUMER, 2017).

Padron (2019) ressalta que o modelo lógico do IFLA LRM reforça a estrutura de relacionamentos dos dados bibliográficos, tornando-a mais adaptada à estrutura dos grafos RDF e favorece a integração dos dados bibliográficos no contexto da Web Semântica. Nesse contexto, destaca-se o modelo geral para agregação de dados.

Figura 23 - Modelo geral para agregações no IFLA LRM



Fonte: Riva, Le Boeuf, Žumer (2017, p. 105, tradução de Padron, 2019, p. 110).

As agregações ocorrem por meio de identificadores persistentes constituídos por conjunto de letras e números ou atribuição de URIs, sendo que a atribuição de URIs é recomendada sempre que possível para estabelecer ligações com vocabulários externos.

O IFLA LRM estabelece três tipos de agregações, sendo 1- agregações com coleções de expressões, neste tipo têm-se uma, duas ou mais expressões que são criadas de maneira independente e publicadas em conjunto em única manifestação; 2 – agregações resultantes de ampliação, este tipo ocorre quando um material se complementa a outro material adicional, como ilustrações, introduções, ou seja, parte complementar ao objeto principal; e 3 – agregações de expressões paralelas, este tipo refere-se à ligação de expressões múltiplas, como o protocolo de pesquisa e o plano de gestão de dados, no contexto dos cadernos de pesquisa.

7 DIRETRIZES SEMÂNTICAS PARA ESTRUTURAÇÃO DE CADERNOS ABERTOS DE PESQUISA

A proposta deste capítulo contempla o último objetivo específico da tese, que consiste na identificação das etapas e elementos para compor um conjunto de diretrizes semânticas para estruturação e, posteriormente, publicação de dados de pesquisa de cadernos de laboratório em formato aberto. Para tanto, buscou-se a compreensão do que sejam diretrizes semânticas e a melhor definição de publicação de dados de pesquisa científica para, então, prosseguir com a definição das etapas e elementos.

Segundo o Michaelis Moderno Dicionário da Língua Portuguesa (2019, on-line), a palavra diretriz pode ser entendida como "linhas gerais que orientam um projeto". De uma maneira geral, pode-se compreender que diretrizes são recomendações para se estabelecer um projeto e atingir um objetivo. Nesta tese, as diretrizes semânticas referem-se ao conjunto de orientações elaboradas com base nos conceitos e tecnologias da Web Semântica e *Linked Data* para estruturação e publicação de cadernos abertos de pesquisa.

A publicação de dados de pesquisa científica segue o conceito de *data publishing* ou *data publication*, correspondendo ao ato de liberar dados de pesquisa em formato aberto para serem usados, modificados e compartilhados por qualquer pessoa para qualquer fim, estando sujeito, no máximo, a requisitos que preservem a proveniência e abertura, conforme orientações da Open Definition (2014) apresentadas no item 4.2 desta tese. Para isso, as diretrizes apresentam um conjunto de tecnologias e melhores práticas necessárias para representação e interoperabilidade das informações anotadas nos cadernos de pesquisa.

Para a publicação de dados, segundo os propósitos desta tese, há de se destacar os principais elementos de uma publicação: a composição dos objetos digitais e as relações conceituais entre eles. A identificação dos objetos digitais e suas composições se dá por meio de mapeamento do ecossistema de dados de pesquisa científica em torno do caderno de pesquisa. Em seguida, é realizada a modelagem dos dados dos objetos digitais para identificar as relações entre as entidades e os atributos para descrevê-los, com vista a atender as tarefas dos usuários.

Após o processo do mapeamento do ecossistema do caderno de pesquisa e modelagem dos dados de pesquisa desses cadernos, prossegue-se com o desenvolvimento das etapas para estruturação e publicação dos dados.

7.1 IDENTIFICAÇÃO DO ECOSSISTEMA DOS CADERNOS DE PESQUISA

Para o Banco Mundial, o ecossistema de dados abertos refere-se à abordagem que trata da questão de dados abertos com enfoque não somente nos dados em si, mas também numa percepção mais ampla que envolve diversos atores, pessoas e organizações, como desenvolvedores e universidades, sejam do setor público ou privado; envolvidos nas dimensões de liderança e aspectos políticos e legais (THE WORLD BANK, 2013).

O ecossistema de dados de pesquisa científica pode ser compreendido como um conjunto de elementos que se relacionam sobre um determinado contexto, como também de atores envolvidos no processo de produção dos dados e maneiras de publicar dados abertos. Um caderno de pesquisa normalmente é temático, isto é, realiza várias pesquisas sobre um determinado assunto. Cada pesquisa gera um conjunto de dados (*datasets*) que pode ser formado por um único elemento ou por um conjunto de elementos relacionados que se complementam com arquivos em planilhas, imagens, anotações técnicas, dentre outros. Esses elementos podem ser gravados em um ou vários arquivos de documentos.

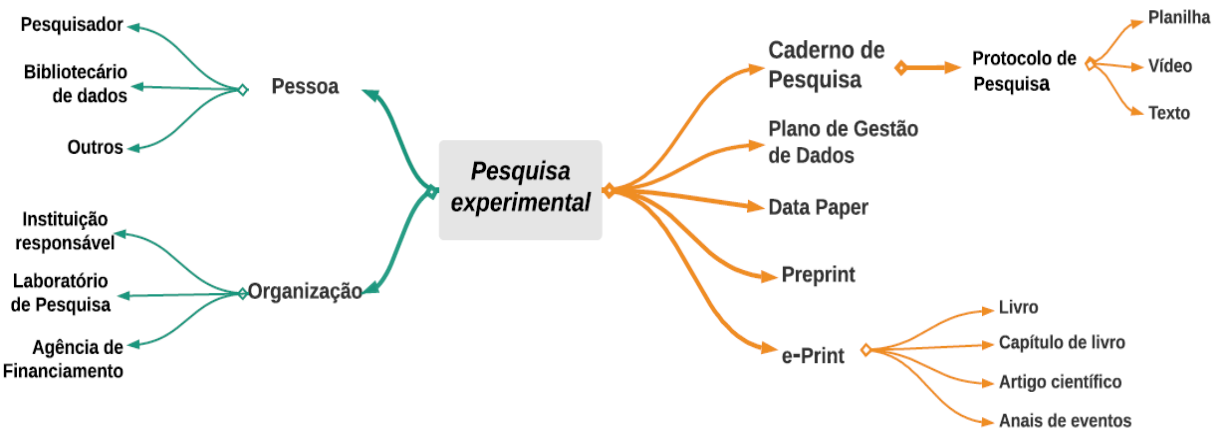
Em uma publicação de dados de pesquisa científica de cadernos de laboratório há o envolvimento de pessoas e organizações, enquanto atores do ecossistema da pesquisa. A pessoa pode ser o pesquisador que desempenha as funções de autor, colaborador, orientador, dentre outras funções de uma pesquisa, além do bibliotecário de dados que facilita a organização e a disponibilização dos dados. A organização pode ser a instituição responsável pela pesquisa, o laboratório que desenvolve a pesquisa e a agência de fomento que financia a pesquisa.

O pesquisador é o principal ator do ecossistema da pesquisa científica no processo de publicação de cadernos de laboratório, tanto como provedor de dados quanto consumidor dos dados publicados em tais cadernos. O pesquisador, na qualidade de provedor de dados, publica um documento contendo os registros ou anotações dos procedimentos do experimento da pesquisa. Esse documento é denominado por Latour e Woolgar (1997) de inscrição literária, no sentido de formalizar literalmente os fenômenos que servirão posteriormente de matéria-prima para a elaboração dos enunciados científicos. Faria-Campos *et al*. (2020, p. 182) o denominam de protocolo, o qual “padroniza um método de laboratório para garantir a replicação bem-sucedida dos resultados por outras pessoas no mesmo ou em outros laboratórios”. Esse mesmo documento é chamado entre pesquisadores e grandes agências de fomento de fluxo de trabalho (*workflows*). Além desses documentos – nomeados de objetos digitais ou recursos digitais - os pesquisadores que são provedores de dados podem publicar

informações complementares que favorecem o enriquecimento dos dados de suas pesquisas, como o Plano de Gestão de Dados (PGD). Nesse sentido, Silva (2019) descreve todo o ciclo de vida dos dados, desde a sua coleta até a documentação completa do processo de pesquisa. Os pesquisadores podem publicar em periódicos ou repositórios de dados os *datas papers* ou artigos de dados que objetivam fornecer conjuntos de dados da pesquisa sem se estender para as inferências do autor. Os pesquisadores podem incluir no processo de publicação os *Preprints*, os quais se referem à versão de um manuscrito antes da avaliação por pares, bem como podem incluir os *e-Prints* que se referem aos documentos já publicados, como livros, capítulos de livros, artigos de periódicos, teses e dissertações etc.

Esse ecossistema pode ser visualizado na taxonomia dos cadernos de pesquisa.

Figura 24 - Taxonomia do Ecossistema de Pesquisa Experimental



Fonte: Elaborado pela autora (2020).

Estes objetos ou recursos digitais podem ser compreendidos da seguinte maneira:

- **Caderno e Protocolo de Pesquisa**

O caderno de pesquisa representado pelo protocolo recebe destaque neste ecossistema por ser o principal objeto de descrição e publicação de uma pesquisa experimental, o qual pode ser apresentado em formato de planilhas, vídeos, textos, dentre outros formatos. O protocolo de pesquisa diz respeito ao registro padronizado dos procedimentos realizados durante uma pesquisa experimental, de modo a garantir a replicação bem sucedida dos resultados (LATOUR; WOOLGAR, 1997). Entende-se que o protocolo de pesquisa seja o principal objeto para a publicação de uma pesquisa científica realizada dentro de um laboratório, em função do detalhamento dos procedimentos realizados, dos equipamentos adotados e dos reagentes utilizados durante a pesquisa, além de declarar os objetivos

pretendidos, a discussão dos dados encontrados na pesquisa e registrar os resultados alcançados, sejam parciais ou finais. Para Latour e Woolgar (1997), os pesquisadores garantem que a inscrição literária, a qual é denominada de protocolo de pesquisa, possua uma descrição completa dos procedimentos, equipamentos, instrumentos químicos e outros suprimentos utilizados, além da descrição dos objetivos do estudo, a ideia para o projeto experimental, a definição para tamanhos de amostra escolhidas, precauções de segurança e como os resultados foram calculados e relatados. Cada protocolo deve ser entregue acompanhado de documentos textuais e planilhas, caso seja necessária para a identificação e interpretação dos procedimentos de materialização dos objetos de estudo como traços, pontos, gráficos, mapas, espectros, fotografias ou números produzidos por aparelhos de manipulação.

A Foster (2018), classifica os cadernos de pesquisa como parte integrante da terceira dimensão *Open Reproducible Research*, da taxonomia da Ciência Aberta, a qual se constitui no ato de oferecer aos usuários livre acesso a elementos experimentais para permitir a reprodução da pesquisa, independente de seus resultados. Vale ressaltar que as condições de reprodutibilidade e repetibilidade são diferentes. A condição de medição da reprodutibilidade inclui diferentes locais, diferentes operadores, diferentes sistemas de medição e medições repetidas no mesmo objeto ou em objetos similares. Enquanto que a condição de repetibilidade inclui os mesmos procedimentos de medição, os mesmos operadores, o mesmo sistema de medição, as mesmas condições de operação e o mesmo local. Partindo desse entendimento, pode-se considerar que o caderno de pesquisa pode proporcionar a reprodutibilidade em diferentes condições de medição e nem sempre conseguirá a repetibilidade em função de fatores participantes do momento. Dessa forma, a descrição dos elementos dos cadernos de pesquisa deve contemplar informações que englobam a realização do experimento, fórmulas e técnicas de medição adotadas. O pesquisador deve informar as variações que interferem no resultado de uma pesquisa, de modo que os consumidores de dados tenham condição de definir entre o estudo a partir da reprodução, replicação ou não replicação da pesquisa científica sem esforços desnecessários (INMETRO, 2012).

Apresenta-se a seguir a compreensão dos outros objetos de pesquisa apresentados na taxonomia do ecossistema da pesquisa experimental, os quais poderão facilitar o desenvolvimento de uma futura publicação ampliada.

- **Plano de Gestão de Dados**

Segundo a Comissão Europeia, o Plano de Gestão de Dados (PGD) é um documento que descreve como os dados de pesquisa compilados ou gerados serão tratados durante um

projeto de pesquisa e, após de concluído, descreve quais dados serão recolhidos, seguindo metodologia e padrões específicos, e como serão compartilhados, se serão abertos e como se realizará a curadoria digital. Segundo Silva (2019, p. 55) o PGD requer uma sequência documentada de ações destinadas a identificar, assegurar recursos, coletar, manter, proteger e utilizar os arquivos de dados. Isto inclui a obtenção de financiamento e a identificação dos recursos técnicos e de pessoal para o completo ciclo de gestão de dados.

De acordo com a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), o plano de gestão de dados é um componente que vem sendo exigido pelas principais agências governamentais ou privadas de financiamento na América do Norte, Europa e Austrália como requisito para análise do financiamento da pesquisa. No Brasil, a FAPESP é uma agência de fomento que solicita a elaboração do PGD e orienta sobre quanto as principais informações devem conter no plano, tais como os dados que serão gerados e como serão preservados e disponibilizados, considerando questões éticas, legais e de confidencialidade.

- ***Data Paper***

É uma publicação que descreve uma coleção de dados no formato em que foram gerados com descrição em metadados estruturados e legível por máquinas e humanos, cuja característica se destaca por não possuírem analises e conclusões de autores.

Para Curty (2017) os artigos de dados ou *data papers* são submetidos a periódicos científicos e repositórios de dados para serem publicados, a partir das etapas:

> [...] primeiramente, os autores selecionam um periódico de dados adequado à pesquisa, e verificam quais repositórios são aceitos/autorizados pelos periódicos. Os autores redigem o artigo de dados de acordo com as instruções, modelos e ferramentas recomendadas pelo periódico. Na segunda etapa, os autores submetem o conjunto de dados ao repositório e recebem um identificador e os metadados do artigo, mas não necessariamente disponibilizam os dados abertamente, podendo deixá-los abertos somente ao editor do periódico, para fins de avaliação pelos pares. Os autores então submetem os artigos de dados ao periódico, adicionando o identificador e os metadados providos pelo repositório no ato do arquivamento. Na terceira etapa, o artigo é submetido ao processo de avaliação e revisão pelos pares, sendo que, uma vez aceito, os dados deverão ser disponibilizados de forma aberta, sem restrições de acesso. (CURTY, 2017, p. 12).

A publicação dos dados de pesquisa científica, a partir de um *data paper* proporciona a facilidade de compartilhamento e a reutilização de dados para fins de novas análises e interpretações, com os devidos créditos aos pesquisadores que os coletaram.

- ***Preprint***

Segundo Cunha e Cavalcanti (2008, p. 290), "*preprints* referem a tiragens antecipadas de um artigo, ou de trabalhos apresentações a reuniões científicas". Na publicação acadêmica, segundo Spinak (2016), um *preprint* é a versão de um manuscrito antes da avaliação por pares que podem aprovar ou não a publicação formal em um periódico. Para Spinak (2016), a versão *preprint* pode ser um texto completo ou uma versão incompleta, porém o mais comum é uma versão final. Nassi-Calò (2020) esclarece que a prática de publicar *preprints* ainda é tímida no mundo todo, mas a postagem ou depósito desses objetos em servidores Web antes da avaliação por pares é uma maneira de comunicar os resultados de uma pesquisa em acesso aberto e proporcionar maior visibilidade e citação dos artigos publicados.

No contexto dos repositórios de *preprints* destaca-se a plataforma *Emerging Research Information* (EmerRI), implementada em 2020, a partir da cooperação entre a Associação Brasileira de Editores Científicos (ABEC) e o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), cujo objetivo é o de prestar serviços às revistas e editores com o propósito de acelerar a disponibilização dos artigos submetidos a suas revistas, especialmente frente à pandemia do Coronavírus. Sobre o repositório EmerRI vale destacar que os textos dispõem de uma licença EmeRi de distribuição exclusiva e uma declaração de revisão preliminar. No entanto, a revisão preliminar não parece ser uma prática comum entre os repositórios de *preprints*, mas sim uma situação emergencial em função da pandemia do Coronavírus.

- ***e-Print***

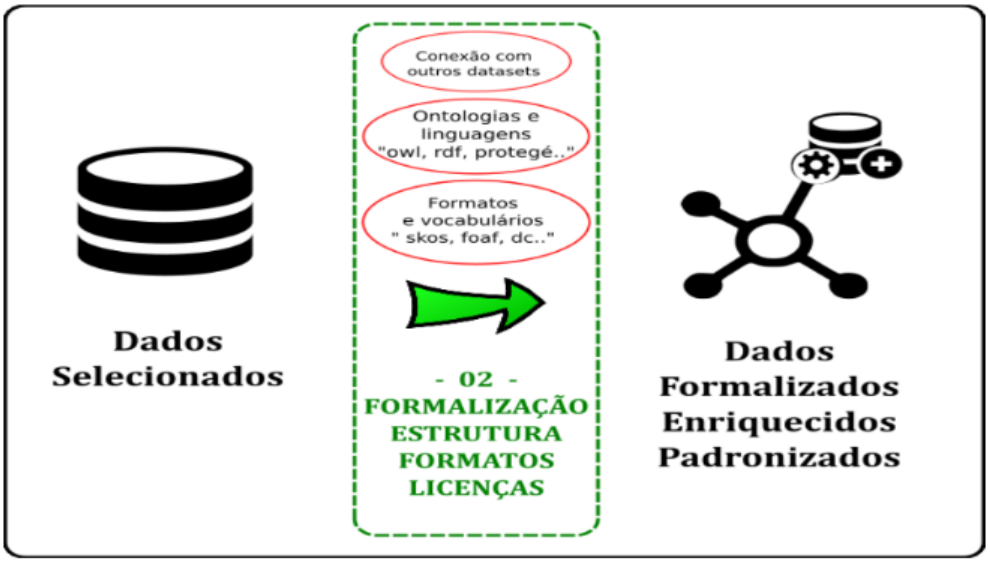
O *e-Print* pode ser compreendido como a configuração final de uma pesquisa científica, o qual apresenta resultados de pesquisas científicas e são publicados após avaliação por pares, como é o caso de um artigo científico publicado em periódicos ou em anais de uma comunicação científica. De acordo com Verhaar (2008, p. 11), os *e-Prints* são compreendidos como "um recurso textual como trabalho acadêmico original, que se destina a ser lido por pessoas, que apresenta algumas reinvindicações acadêmicas e que geralmente contém uma interpretação ou uma análise de determinados dados primários". Sales (2015) complementa que e-*Prints* são considerados objetos de pesquisa em formato digital usados para comunicar resultados de atividades de pesquisa acadêmica, como artigos de periódicos, livros, teses e dissertações, dentre outros.

7.2 ETAPAS PARA ESTRUTURAÇÃO E PUBLICAÇÃO DE DADOS DE PESQUISA CIENTÍFICA DE CADERNOS ABERTOS DE PESQUISA

Para publicar dados de pesquisa científica registrados em cadernos abertos de pesquisa, adotará a fase 2 – Formalização, Estrutura, Formatos e Licenças do modelo proposto por Santarem Segundo (2018), denominado ‘Fluxo Organizacional para Publicação de Dados’. Trata-se de um modelo segmentado em fases, que organiza o caminho por qual um projeto de publicação de dados deve percorrer. É um modelo amplo, ou seja, não se destina a uma área específica. Sendo assim, optou por seguir as orientações da etapa 2 para a estruturação de dados para fins de publicação.

A ação de estruturar dados para efeito de publicação tem início com os dados e fonte de dados previamente selecionados. Depois disso, segue com a etapa de estruturação dos dados, a qual é estabelecida por Santarem Segundo (2018) como a etapa em que se atribuem aos dados características técnicas que os transformam em semânticos. Nesse contexto, a atribuição de formatos, vocabulários padronizados e linguagens é uma ação necessária para que os dados selecionados possam ser enriquecidos e publicados de forma aberta e semântica. Destaca-se que o estudo dos modelos de ciclo de vida dos dados, princípios FAIR, melhores práticas do W3C e as discussões sobre as tecnologias da Web Semântica, possibilitaram a definição de um conjunto de tecnologias, formatos, vocabulários e licenças que permitem a estruturação e publicação de dados de pesquisa de cadernos de laboratório nas condições favoráveis a sua reutilização, redistribuição e conexão a outros conjuntos de dados.

Figura 25 - Formalização, Estruturação, Formatos e Licença



Fonte: Santarem Segundo (2018, p. 128).

Uma publicação requer a organização dos dados e uma das tarefas de organização é a padronização de tais dados, ou seja, usar os mesmos tipos de informações, vocabulários controlados, associar termos que tenham conceitos correspondentes ou similares e atribuir os vocabulários apropriados, o que inclui classes e propriedades.

Para a composição das diretrizes semânticas foram definidas as seguintes etapas para a estruturação dos dados de pesquisa de cadernos de laboratório:

1 – Modelagem dos conjuntos de dados dos cadernos abertos de pesquisa;

2 – Mapeamentos e descrição de metadados;

3 – Descrição dos vocabulários selecionados;

4 – Descrição dos vocabulários utilizados para enriquecimento de dados;

5 – Definição de licenças de uso;

6 – Mapeamento das propriedades de vocabulários;

7 – Mapeamento de propriedades para relacionamento.

Após a definição das etapas de estruturação de dados, foi realizada uma análise dos elementos quanto ao alcance dos dados serem encontráveis, acessíveis, interoperáveis e reutilizáveis, a partir da aplicação dos princípios FAIR, das tecnologias da Web Semântica e dos conceitos do *Linked Data*.

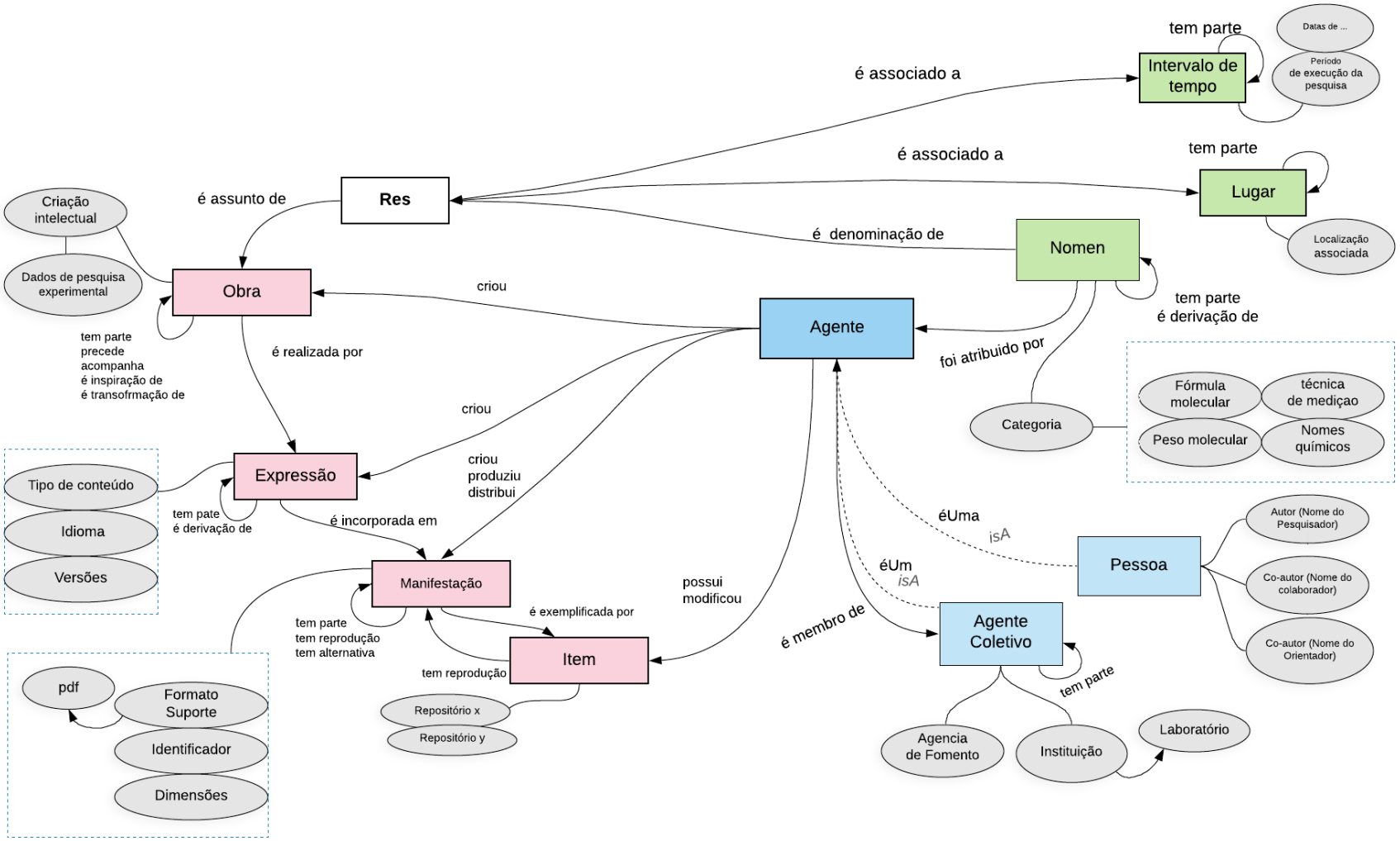
7.2.1 Modelagem dos Conjuntos de Dados dos Cadernos Abertos de Pesquisa

A modelagem para o propósito desta tese consiste em descrever formalmente os vários aspectos dos dados de pesquisa em torno dos cadernos de pesquisa e explicitar os relacionamentos entre eles, bem como identificar os elementos necessários para a descrição com base nas tarefas desempenhadas pelos usuários as quais são encontrar, identificar, selecionar, obter e explorar, conforme descritos na seção 6.3.5. A proposta do modelo conceitual IFLA LRM, foi escolhida por ser um modelo do tipo entidade-relacionamento de alto nível, que busca contemplar todos os aspectos do universo bibliográfico e adotar a forma necessária para sua utilização em aplicações de dados abertos ligados. Destaca, ainda, que o modelo IFLA LRM adota termos e definições aplicáveis de maneira genérica a todos os tipos de recursos e não apenas em operações técnicas de bibliotecas. Dessa forma, considera-se adequado o uso das entidades que são o foco de interesse do usuário de um sistema de informação bibliográfico, bem como os atributos que descrevem e caracterizam as instâncias das entidades com mais detalhes, e as relações entre as entidades.

Para ser considerada uma implementação IFLA LRM as relações estruturais entre as entidades obra, expressão, manifestação e item são essenciais ao modelo. No entanto, os atributos e outros relacionamentos declarados no modelo não são necessários para a implementação, podendo omitir ou adaptar para aplicações específicas (IFLA LRM, 2017). Apesar de o modelo permitir adaptações em suas aplicações, neste estudo manteve-se as entidades originais do modelo, mas inseriu-se atributos relacionados com as especificidades das pesquisas experimentais realizadas em laboratórios.

A figura 26, a seguir apresenta as partes dos objetos digitais definidas pelas entidades, classes de nível 1, 2 e 3, vinculadas ao relacionamento e à rotulação dos participantes dessa relação, de modo a associar a cada um o papel que o mesmo desempenha.

Figura 26 – Modelagem de Dados de Cadernos de Pesquisa - Modelo IFLA LRM



Fonte: Adaptação de IFLA (2017).

Nas diretrizes, a entidade ***Res*** refere-se ao recurso descrito. A *Res*, a superclasse do modelo, única classe de primeiro nível, referenciada como LRM-E1, representa o assunto da obra. A entidade *Res* possui o relacionamento de associação e os atributos podem ser por categoria, por exemplo, objeto, obra, conceito, evento, família e entidade corporativa; e ainda, por nota, ou seja, qualquer atributo sobre *Res* que não esteja registrado através do uso de atributos e/ou relacionamentos específicos é utilizada uma nota geral para informar.

As designações de relacionamento de *Res* são:

Res <é assunto de> Obra
Res <tem denominação> Nomen
Res <é associado a> Lugar
Res <é associado a> Intervalo de tempo
Res <é associado a> Res

Considerando que a entidade *Res* tem denominação *Nomen*, os atributos para assunto estão listados na entidade *Nomen*.

As entidades obra, expressão, manifestação e item, com destaque na cor rosa, são essenciais ao modelo e, portanto precisam ser indicadas. Estas entidades representam o grupo 1 do modelo conceitual FRBR.

A entidade **obra**, indicada em LRM-E2, é a criação do conteúdo intelectual em si, independente do seu suporte. A nota de escopo da entidade descreve que a conexão lógica entre uma obra e o agente relacionado serve como base tanto para identificar um agente responsável por uma obra individual, quanto para garantir que todas as obras de um agente em particular estejam vinculadas com tal agente. Os relacionamentos designados são:

Obra <tem assunto> Res
Obra <tem parte> obra
Obra <precede> obra
Obra <acompanha> obra
Obra <é inspiração> obra
Obra <é transformação de> obra
Obra < é realizada por>expressão
Obra <é criada por> Agente

Os atributos podem caracterizar a obra quanto à intenção de término (monografia, seriado), domínio criativo (literatura, música, artes plásticas) e forma ou gênero (romance, poemas, concertos, pintura, fotografia). No caso dos dados de pesquisa científica de cadernos de laboratório, considera-se que a forma da obra seja ‘dados de pesquisa experimental’.

Uma **expressão**, indicada em LRM-E3, é uma combinação distinta de sinais identificáveis de qualquer forma ou natureza (incluindo sinais visuais, auditivos ou gestuais) destinados a transmitir conteúdo intelectual ou artístico. A expressão é mais facilmente

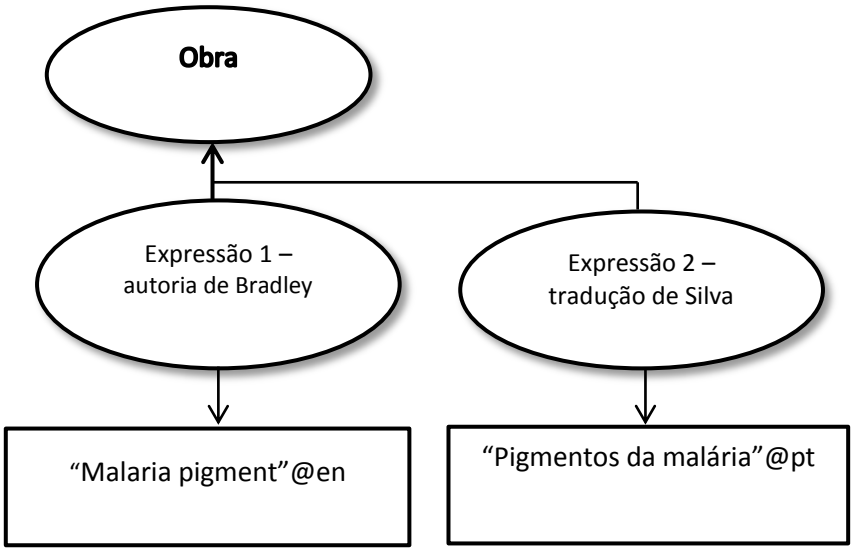
identificada pelas versões e idiomas do conteúdo pelos quais são criados. Por exemplo, a pesquisa científica pode ser reproduzida e publicada no idioma original e traduzida em outro idioma.

A entidade expressão apresenta os seguintes relacionamentos

Expressão <foi criada por> agente
Expressão <tem parte> expressão
Expressão <é derivada de> expressão
Expressão <é incorporada em> manifestação

Para dados de pesquisa será mais comum indicar o idioma da versão original, por exemplo, português (pt), inglês (en) e assim por diante. Ao ser traduzida para outro idioma passa a ter a sua segunda expressão.

Figura 27 - Exemplo de duas expressões de uma obra



Fonte: Elaborada pela autora (2020).

O mesmo raciocínio pode ocorrer para outros objetos digitais do ecossistema dos cadernos de pesquisa. Considerando um objeto *e-Print* (livro, artigo etc.) as expressões são mais facilmente reconhecidas, como é o caso das edições, por exemplo.

A **manifestação**, indicada em LRM-E4, é a materialização de uma expressão de uma obra, ou seja, a representação de todos os objetos físicos que possuem as mesmas características relacionadas ao conteúdo intelectual como forma física. Os atributos da manifestação referem-se às categorias quanto ao suporte geral (exemplo: folha de papel), material físico usado na produção de suportes (exemplo: plástico), material físico aplicado ao material base dos suportes (por exemplo: emulsão química, tinta a óleo [aplicada à tela]),

meios utilizados para registrar a anotação (analógico, digital, acústico). Nesse entendimento, pode-se dizer que a manifestação dos cadernos abertos de pesquisa seja o suporte digital. Além desses atributos da manifestação, tem-se a extensão do documento (número de páginas, polegadas, diâmetro, a dimensão do documento), o público alvo destinado pela manifestação (por exemplo, uma manifestação pode ter sido criada para um público com deficiência visual), as condições de acesso (requisitos do sistema, modo de acesso, etc) e direitos de uso (apresentação de uma licença de uso).

A entidade **item**, indicada em LRM-E5, exemplifica a manifestação e normalmente configura-se como um objeto físico único ou em vários objetos físicos. Pode-se exemplificar que o item seja um exemplar ou vários exemplares que se encontram armazenados em um acervo de biblioteca, um arquivo pessoal, em uma plataforma digital, dentre outros ambientes informacionais.

Os relacionamentos designados para o item são:

Item <exemplifica> manifestação
Item <tem reprodução> manifestação
Item<é modificado por> agente
Item <é propriedade de> agente

O protocolo de pesquisa ou *Workflows* poderá ser armazenado em um repositório e a sua tradução em outro. Assim, o artigo científico poderá ser armazenado em um repositório e a versão em outro idioma em outro.

A entidade **agente**, indicada em LRM-E6, classe de segundo nível, se desmembra nas entidades pessoa e agente coletivo, classes de terceiro nível. As entidades com destaque na cor azul representam o grupo 2 do FRBR e entidades do modelo FRAD.

A entidade **pessoa**, indicada em LRM E7, corresponde ao indivíduo responsável pela criação intelectual do conjunto de dados. Neste estudo, considera-se que os agentes pessoais sejam todos os pesquisadores com os seus diferentes níveis de participação na pesquisa, por exemplo, autor principal, autor colaborador e autor orientador. Enquanto que para os *e-Prints*, os agentes pessoais são classificados em autores e coautores. No caso de livros, faz-se necessário considerar ilustradores, tradutores e demais pessoas envolvidas no processo intelectual.

A entidade **agente coletivo**, indicada em LRM E8, corresponde a uma reunião ou organização de pessoas com um nome particular e capaz de atuar como unidade (IFLA, 2017).

No cenário dos dados de pesquisa científica de cadernos de laboratório podem-se estabelecer as seguintes designações de relacionamento:

Agente <é uma> Pessoa
Pessoa <é um> Autor
Pessoa <é um> Colaborador
Pessoa <é um> Orientador
Agente <é um>Agente Coletivo
Agente Coletivo <é uma>Instituição
Agente <é membro de> Agente Coletivo
Pessoa < é membro de> Instituição
Autor < é membro de> Instituição
Colaborador < é membro de > Instituição
Orientador < é membro de> Instituição
Agente Coletivo <tem parte> Agente Coletivo
Instituição <tem parte> Laboratório
Agência de Fomento < é um>Agente coletivo

Nota-se que as subclasses foram desmembradas para identificar os atributos que descrevem os principais atores do ecossistema da pesquisa científica de cadernos de laboratório. Os atributos para entidade agente são: informação de contato, campo de atividade, idioma e profissão/ocupação. Esta última é indicada para pessoas. O agente coletivo não possui nenhum atributo restrito. Além dessas, considerou-se importante acrescentar na entidade pessoa os atributos referentes a datas – data de nascimento e morte, quando for o caso -, e destacar as instituições e departamentos pelos quais os autores são membros.

As entidades *nomen*, lugar e intervalo de tempo, com destaque na cor verde, referem-se ao grupo 3 do modelo conceitual FRBR e entidades do modelo conceitual FRSAD.

A entidade ***nomen***, indicado em E9, refere-se à associação entre uma entidade e uma designação referente a ela. Os principais atributos dos conjuntos de dados dos cadernos de pesquisa encontram-se destacados na entidade *nomen*. As principais designações de relacionamento da entidade *nomen* referem-se a

Nomen <é denominação de>Res -----> Res é o assunto da obra
Nomen <foi atribuído por>Agente
Nomen <tem derivação> Nomen
Nomen <tem parte> Nomen

A entidade *nomen*, permite a indicação de atributos relacionados aos conceitos e especificidades das pesquisas experimentais. Os atributos dessa entidade são classificados em categoria, um tipo ao qual o *nomen* pertence (coisa nomeada: nome de pessoa, cargo), função do *nomen* indicado por meio de identificadores; *string*, a combinação de sinais dentro de um

sistema de escrita, símbolos de estrutura química, ou qualquer outro tipo de sinal; em esquema incluem esquemas de codificação de valores (exemplo, listas de autoridades de assunto), esquemas de codificação de sintaxe (exemplo, padrões para codificação de datas); público que se pretende atender; contexto de uso a que se destina a pesquisa; fonte de referência; idioma; e script.

A entidade **lugar**, indicada em E10, corresponde a uma extensão de espaço. Em notas de escopo consta que os lugares são geralmente identificados através de um objeto físico (uma característica geográfica ou um objeto feito pelo homem) ou devido à sua relevância em relação a um determinado agente (entidades geopolíticas como países e cidades) ou como a localização de um evento. Os atributos são classificados em categoria: cidade, país e continente expresso no idioma; localização, enquanto delimitação do território físico do lugar. As principais designações de relacionamento da entidade Lugar referem-se a:

Lugar <tem parte> Lugar
Lugar <é parte de> Lugar
Lugar <é associação com> Res
Lugar <tem associação com> Res

A entidade **intervalo de tempo**, indicada em E11, corresponde a uma extensão temporal que possui um inicio, um fim e uma duração. As principais designações de relacionamento da entidade intervalo de tempo correspondem a:

Intervalo de tempo <tem parte> Intervalo de tempo
Intervalo de tempo <é parte de> Intervalo de tempo
Intervalo de tempo <é associado com>Res
Intervalo de tempo <tem associado com>Res

Este estudo sobre modelo conceitual IFLA LRM não pretendeu ser exaustivo, mas destacar as entidades, relacionamento e atributos que podem ser pertinentes à descrição dos cadernos abertos de pesquisa. Dito isso, evidencia-se que a modelagem proporcionou o refinamento dos metadados, os quais serão compilados e descritos nas seções seguintes.

7.2.2 Mapeamento de Metadados

O mapeamento de metadados identificou os vários aspectos dos dados que descrevem e individualizam os objetos que compõem o ecossistema da pesquisa experimental, inclusive do caderno de pesquisa constituido por protocolos. O delineamento dos metadados ocorreu por meio da análise das características dos dados publicados pelos cadernos *UsefulChem*, *LabScribbles* e *Openlabnotebooks*, pela análise da literatura e, principalmente, do alinhamento dos conceitos do modelo IFLA LRM ao contexto dos cadernos de pesquisa, os quais permitiram identificar as classes, propriedades e atributos dos dados dos objetos em estudo.

Os metadados identificados para os tipos de objetos digitais são descritos a seguir.

Quadro 15 - Mapeamento de metadados dos objetos digitais

Protocolo	**Plano de Gestão de Dados**	***Preprint e Data Paper***	***e-Print***[25]
Identificador do registro	Identificador do registro	Identificador do registro	Identificador do registro
Data e horário do registro	Data e horário do registro	Data e horário do registro	Data e horário do registro
-	-	-	ISSN
Nome do pesquisador	Nome do pesquisador	Autor	Autor
Nome dos pesquisadores colaboradores	Nome dos pesquisadores colaboradores	Co-autores	Co-autores
Data de nascimento	Data de nascimento	Data de nascimento	Data de nascimento
Data de morte	Data de morte	Data de morte	Data de morte
Profissão/Ocupação	Profissão/Ocupação	Profissão/Ocupação	Profissão/Ocupação
Instituição vinculada	Instituição vinculada	Instituição vinculada	Instituição vinculada
Departamento da Instituição vinculada	Departamento da Instituição vinculada	Departamento da Instituição vinculada	Departamento da Instituição vinculada
Patrocinador da pesquisa	Patrocinador da pesquisa	Financiador	Financiador
Campo de atividade dos agentes	Campo de atividade dos agentes	Campo de atividade dos agentes	Campo de atividade do autor
Instituição responsável pela preservação dos dados	Instituição responsável pela preservação dos dados	-	-
Idioma do pesquisador	Idioma do pesquisador	Idioma do pesquisador	Idioma do pesquisador
Informação de contato dos agentes	Informação de contato do pesquisador	Informação de contato do autor	Informação de contato do autor
Identificadores dos agentes	Identificadores dos agentes	Identificadores dos agentes	Identificadores dos agentes
-	Número do subsídio	-	-
Nome do projeto	Nome do projeto	Título do *preprint*	Título do artigo
Subtítulo do projeto	Subtítulo do projeto	Subtítulo do *preprint*	Subtítulo do artigo

[25] Para cada tipo de *e-Print* há um conjunto de atributos, nesse exemplo considera um artigo científico publicado em periódico científico.

-	-	-	Título do periódico
Idioma da pesquisa	Idioma da pesquisa	Idioma da pesquisa	Idioma da pesquisa
Formato	Formato	Formato	Formato
Tipo	Tipo	Tipo	Tipo
Tamanho do arquivo	Tamanho do arquivo	-	-
Software necessário	Software necessário	-	-
-	-	-	Designação numérica e/ou alfabética
Local da pesquisa	Local de pesquisa	-	Local de publicação
-	-	-	Editora de publicação
Data da primeira versão do projeto	Data da primeira versão do projeto	-	-
Data da última versão	Data da última versão	Data de produção	Ano de publicação
Período de execução da pesquisa	Período de execução da pesquisa	-	-
-	-	Data de submissão ao periódico/conferência	Data de submissão ao periódico
-	-	-	Data do aceite
-	-	-	Página inicial
-	-	-	Pagina final
Número de páginas	Número de páginas	Número de páginas	Número de páginas
Público	Público	-	-
Objetivo pretendido	Objetivo pretendido	-	-
Reagentes	-	-	-
Equipamentos	-	-	-
Descrição dos procedimentos	Descrição da pesquisa	-	-
Fórmula molecular	-	-	-
Peso molecular	-	-	-
Técnica de medição	-	-	-
Nomes químicos	-	-	-
Nome comercial	-	-	-
Resultados alcançados	Resumo	Resumo	Resumo
Palavras-chave	Palavras-chave	Palavras-chave	Palavras-chave
Fonte de dados	Fonte de dados	Fonte de dados	Fonte de dados
Declaração de proveniência	Declaração de proveniência	Declaração de proveniência	Declaração de proveniência
Licença de uso	Licença de uso	Licença de uso	Licença de uso
Titular dos direitos	Titular dos direitos	Titular dos direitos	*Copyrights*

Fonte: Elaborado pela autora (2020).

O mapeamento do Protocolo de Pesquisa, principal objeto do ecossistema dos cadernos de pesquisa, foi realizado a partir de consultas em cadernos temáticos e na revisão da literatura. Os metadados extraídos do caderno *UsefulChem* foram objetivo do experimento, descrição dos procedimentos realizados, resultados alcançados, discussão e histórico de

modificações do experimento. Em análise aos cadernos *LabScribbles e Openlabnotebooks*, extraíram-se os metadados referentes ao direito de acesso, formato do arquivo e tipo do documento. Os metadados reagentes, equipamentos, fórmula molecular, peso molecular, técnica de medição e nomes químicos e comerciais foram estabelecidos a partir das necessidades dos pesquisadores, expressas na revisão de literatura, principalmente na obra de Latour e Woolgar (1997) e Faria-Campos *et al*. (2020). O refinamento desses metadados foi realizado a partir dos relacionamentos e atributos sugeridos no modelo IFLA LRM (2017).

Os metadados referentes ao Plano de Gestão de Dados foram extraídos das obras de Sayão e Sales (2015) e Silva (2019), além de consultas realizadas em agências de fomento. Estas últimas apenas confirmaram as informações dos autores mencionados. As informações sobre os agentes e proveniência dos dados foram elaboradas a partir do IFLA LRM (2017).

Para mapear os metadados dos *Preprints*, baseou-se na definição de documento ainda não publicado e nas orientações de submissão do repositório *Emerging Research Information* (EmeRi). Para os *Preprints* e *Data Papers* definiu-se um conjunto de elementos essenciais, considerados mínimos para qualquer tipo de descrição. As informações sobre os agentes e proveniência dos dados foram elaboradas a partir do modelo conceitual IFLA LRM (2017).

No que se trata dos objetos de *e-Print*, definiu-se adotar como exemplo, nestas diretrizes, o artigo científico por ter mais elementos a serem trabalhados na estruturação. Os metadados foram identificados a partir do mapeamento dos elementos estabelecidos pela *International Standard Bibliographic Description* (ISBD), atributos do modelo IFLA LRM e consultas realizadas no código de catalogação *Resource Description & Access* (RDA).

Após o mapeamento dos metadados dos objetos digitais que compõem o ecossistema da pesquisa experimental, realizou-se a descrição dos metadados a partir dos sites do DCMI e Schema.org, além do código RDA. Apesar do modelo conceitual IFLA LRM não prever todos os tipos de metadados estabelecidos no quadro 16, atentou-se pelas orientações das melhores práticas do Consórcio W3C, dos princípios FAIR e dos modelos de ciclo de vida dos dados estudados nesta tese. Os metadados a seguir estão organizados em ordem alfabética para facilitar a encontrabilidade.

Quadro 16 - Descrição dos Metadados do Conjunto de Dados de Caderno de Pesquisa

Metadados	Descrição
Agência de Fomento	Uma pessoa ou organização que apoia (patrocina) algo por meio de algum tipo de contribuição financeira.
Assunto	Normalmente, o assunto será representado usando palavras-chave. No entanto, a melhor prática é se referir ao assunto adotando um URI e se isso não for possível, um valor literal que identifique o assunto pode ser fornecido. A recomendação é usar vocabulário controlado, sempre que possível, entre os exemplos encontram-se o Pubchem, o LCSH e o MeSH. *Vide* a *tag* identificador de assunto, ponto de acesso autorizado e ponto de acesso variante.
Assunto mais amplo	Classificação do assunto a partir de termos mais gerais. Ao mencionar o termo geral deve-se informar o termo específico. Na ausência da classificação do termo, adote a *tag* assunto. A prática recomendada é adotar vocabulário controlado sempre que possível. *Vide* exemplos na *tag* assunto e identificador de assunto.
Assunto mais específico	Classificação do assunto a partir de termos mais específicos. Ao mencionar o termo específico deve-se informar o termo geral. Na ausência da classificação do termo, adote a *tag* assunto. A prática recomendada é adotar vocabulário controlado sempre que possível. *Vide* exemplos na *tag* assunto e identificador de assunto.
Autor/criador	Um agente responsável por criar o recurso. Exemplos de um criador incluem uma pessoa, uma organização ou um serviço. Normalmente, o nome de um criador deve ser usado para indicar a entidade.
Campo de atividade	Um campo de atuação, área de especialização em que o agente está ou esteve envolvido.
Cobertura espacial	Características espaciais do recurso. O vocabulário GeoNames disponibiliza URI adequada a cada local do mundo.
Contribuinte /Colaborador	Uma pessoa ou uma instituição responsável por fazer contribuições para o recurso. A recomendação para o uso de nomes de pessoas ou organizações como criadores também se aplicam aos colaboradores. Normalmente, o nome de um colaborador deve ser usado para indicar a entidade.
Criado por	Usuário que realizou o registro dos objetos digitais na plataforma digital. Esta *tag* trata-se de um metadado administrativo e é gerado automaticamente.
Data	Período de tempo associado à pesquisa. A prática recomendada é expressar a data de acordo com a ISO 8601-1. Informar o tempo de execução da pesquisa.
Data de aceite	A data em que o recurso foi aceito. Normalmente essa *tag* é importante para teses e dissertações, artigos e comunicações científicas submetidas a bancas, periódicos e eventos sucessivamente.
Data de criação	Data original de criação e horário de inserção dos dados em uma plataforma digital.
Data de direitos autorais	Data dos direitos autorais do recurso. Normalmente se indica o ano.
Data de modificação	Data em que o recurso foi modificado. Esta *tag* será preenchida automaticamente quando se tratar de metadados administrativos, ou seja, se estiver tratando da modificação do registro na plataforma digital. Se estiver tratando da modificação da pesquisa, a modificação deverá ser feita pelo(s) autor(es) .
Data de morte	Data de falecimento do(s) autor(es).
Data de nascimento	Data de nascimento do(s) autor(es).
Data de produção	Data de produção é destinada para recursos não publicados, por exemplo, teses e dissertações.
Data de publicação	Data de publicação de um recurso.
Data de submissão	Data de envio do recurso a um periódico ou evento para fins de avaliação por pares.
Declaração de direitos	Fornecer uma declaração sobre os direitos intelectuais de um recurso ou um

	documento legal que concede permissão oficial para fazer algo com um recurso.
Declaração de proveniência	Uma declaração de quaisquer alterações na propriedade e custódia de um recurso, desde a sua criação e que sejam significativas por sua autenticidade, integridade e interpretação.
Departamento da Instituição	Refere-se ao departamento ao qual o(s) autor(es) pertencem ou por qual estão realizado a pesquisa. No caso dos cadernos abertos de pesquisa é importante indicar o nome(s) do(s) laboratório(s) pela(s) qual(is) a pesquisa está sendo realizada.
Descrição	A descrição pode incluir, mas não se limita a: um resumo, um índice, uma discussão, uma representação gráfica ou uma conta de texto livre do objeto (recurso) descrito. Pode-se considerar a descrição dos procedimentos adotados no desenvolvimento da pesquisa.
Designação numérica e/ou alfabética	Identifica o volume de publicação ou trabalho de várias partes. Normalmente utilizada por periódicos na publicação de artigos científicos.
Elementos químicos	Os elementos químicos adotados nos experimentos são descritos adotando *strings* textuais e URI em conformidade com um sistema de informação. O vocabulário PubChem traz identificador apropriado a elementos químicos.
Equipamentos	Lista dos equipamentos adotados para realização da pesquisa.
Error	Para indicar ações com falha, mais informações sobre a causa da falha.
Fonte	Um recurso relacionado do qual o recurso descrito é derivado. Esta propriedade deve ser usada com valores não literais. O recurso descrito pode ser derivado do recurso relacionado, no todo ou em parte. A recomendação é identificar o recurso relacionado por meio de um URI ou uma *string* em conformidade com um sistema formal de identificação.
Formato	O formato do arquivo, meio físico ou dimensões do objeto. A prática recomendada é usar vocabulário controlado, quando disponível.
Fórmula molecular	A fórmula molecular, na química, por exemplo, é aquela que descreve o número de átomos numa molécula. A prática recomendada é utilizar *string* textual ou URI em conformidade com um sistema de identificação. O vocabulário PubChem disponibiliza identificadores persistentes para fórmulas moleculares.
Identificador	Uma referência inequívoca ao recurso dentro de um determinado contexto. A prática recomendada é identificar o recurso por meio de *string* textual em conformidade com um sistema de identificação ou URI. Esta propriedade representa qualquer tipo de identificador para qualquer tipo de coisa, como ISBN, ISSN, DOI, ORCID, VIAF, LCSH, MeSH etc. *Vide* identificador de assunto e identificador de pessoa.
Identificador de assunto	Sequência de caracteres associados exclusivamente com o nome, termo, código etc., representando uma obra ou expressão específica. Entre os exemplos incluem: PubChem, LCSH e MeSH.
Identificador de pessoa	Sequência de caracteres associados exclusivamente com uma pessoa, que serve para fazer a diferença dessa pessoa de outras. Entre os exemplos incluem: ORCID e VIAF. *Vide* informação da propriedade Identificador.
Idioma	O idioma do objeto. A prática recomendada é usar um valor não literal que represente um idioma de um vocabulário controlado como a ISO 639-1, ISO 639-2 ou ISO 639-3. O vocabulário Schema.org recomenda ainda seguir um dos códigos de idioma do padrão IETF BCP47.
Informação de contato	Indicar o meio de comunicação com os agentes, seja pessoal ou coletivo. O valor dessa tag pode ser endereço, telefone e e-mail.
Instituição	A instituição é um agente coletivo identificado por um nome corporativo ou coletivo. Esta entidade deve ser informada quando a instituição for autora do conteúdo. Entre os exemplos incluem: ORCID e VIAF. *Vide* informação da propriedade Identificador.
Instituição vinculada	Refere-se à instituição a qual o(s) autor(es) são vinculados. A prática recomendada é identificar a instituição adotando *string* textual ou URI em conformidade com um sistema de identificação.
Licença	Um documento legal dando permissão para fazer algo em relação ao recurso (objeto). Normalmente indicado por URI.

Local de publicação	Indicação que identifica o lugar ou lugares de publicação, editor ou editores. A prática recomendada é utilizar *string* textual ou URI em conformidade com um sistema de identificação. O vocabulário GeoNames disponibiliza URI adequada a cada local do mundo.
Logs de acesso	O Log de acesso é uma *tag* automática que registra todas as modificações no registro dos recursos.
Nome do editor	Nome de uma pessoa, família ou agente coletivo responsável pela publicação, lançamento ou edição de um recurso. Esta propriedade é apropriada para *e-prints* como livros e periódicos.
Número do subsídio	Número do subsidio recebido como apoio financeiro de alguma agência de fomento.
Número total de páginas	Informar o total do número de páginas de um recurso, por exemplo, 120 p. para documentos publicados e 120 f. para documentos não publicados.
Objetivo pretendido	Descrever os objetivos propostos que se busca alcançar
Outra informação do título	Informação que aparece em conjunto e é subordina ao título principal de um recurso.
Página final	A página em que o trabalho finaliza. Esta *tag* é recomendada para artigos científicos e capítulos de livros que se encontram entre muitos outros textos diversos.
Página inicial	A página em que o trabalho começa. Esta *tag* é recomendada para artigos científicos e capítulos de livros que se encontram entre muitos outros textos diversos.
Palavras-chave	Palavras-chave ou *tags* usadas para descrever o conteúdo do recurso.
Período de encerramento da pesquisa	O tempo final de algo. Para ações que abrangem um período de tempo, quando a ação foi executada. Para mídia, incluindo áudio e vídeo, é o intervalo de tempo do final de um clipe em um arquivo maior.
Período de execução da pesquisa	Informar data ou período de inicio de fim de realização da pesquisa. Para datas e horas a prática recomendada é seguir o formato ISO 8601.
Peso molecular	O peso da(s) molécula(s). A prática recomendada é adotar um vocabulário controlado. O PubChem apresenta identificador que representa pesos moleculares.
Ponto de acesso autorizado	O ponto de acesso pode ser: nome, termo, código etc., representando uma obra ou expressão específica. O ponto de acesso autorizado é o ponto de acesso preferido para representar uma entidade e construído de acordo com regras e padrões.
Ponto de acesso variante	É uma forma do nome que não a escolhida como ponto de acesso autorizado para a entidade. *Vide* descrição para ponto de acesso autorizado.
Procedimentos	Listas os procedimentos adotados durante a pesquisa.
Profissão/Ocupação	Uma profissão ou ocupação em que a pessoa trabalha.
Público	Público-alvo, ou seja, um grupo para o qual algo foi criado.
Reagentes	Lista de reagentes adotadas no experimento.
Resultados alcançados	O resultado produzido na ação.
Status da ação	Indica a disposição atual da ação. A prática recomendada é indicar uma *string* textual informando: em andamento ou encerrada.
Software necessário	Dispositivo necessário para executar os aplicativos utilizados na pesquisa.
Subtítulo	O subtituto é uma parte que complementa o título principal. Outro título ou título alternativo.
Tamanho do arquivo	Registro do tamanho do arquivo ou aplicativo, por exemplo 18MB.
Técnica de medição	Uma técnica ou tecnologia usada em um *dataset* correspondendo ao método usado para medir a(s) variável (is) correspondente(s). Isso é orientado para a publicação de conjuntos de dados científicos e acadêmicos, mas pode ter uma aplicabilidade mais ampla; não pretende ser uma representação completa da medição, mas sim um resumo de alto nível para a descoberta do conjunto de dados.
Tipo	A natureza ou gênero do recurso. A prática recomendada é usar um vocabulário controlado.
Tipo de interação do usuário	A ação que representa o tipo de interação. Essa ação é utilizada para medir a satisfação do usuário quanto ao recurso.
Titular dos direitos	Uma pessoa ou organização que possui ou gerencia direitos sobre o recurso.

	A prática recomendada é se referir ao detentor dos direitos com um URI. Se isso não for possível, pode ser fornecido um valor literal que identifique o detentor dos direitos.
Título	Título fornecido ao objeto ou conjunto de dados. O título pode ser uma palavra, caractere ou grupo de palavras e/ou caracteres que nomeiam um recurso (objeto) ou uma obra nele contida.

Fonte: Elaborado pela autora (2020).

Estes elementos podem ser classificadas em tipos de metadados como administrativos, descritivos, preservação, proveniência e uso. Os metadados administrativos são utilizados no gerenciamento e administração do registro do objeto, os quais geram automaticamente as informações sobre o dia, o horário e o nome da pessoa que inseriu o objeto no sistema e as modificações feitas nele. As principais *tags* são: data de criação, data de modificação, horário, quem criou e histórico de acesso.

Os metadados descritivos são adotados para descrever e representar as informações do objeto, esses elementos são preenchidos com base no recurso descrito. Para a publicação de protocolos de pesquisa, selecionou-se metadados que individualizem uma pesquisa de outra e que favoreçam a sua recuperação, como nomes dos agentes e seus atributos, nome do projeto, título dos *preprints*, *data papers* e *e-prints*, idiomas, formatos, tipos, designação numérica e/ou alfabética (caso específico dos artigos científicos), local da pesquisa e da publicação, datas, números de páginas, público pretendido, reagentes, descrição dos procedimentos, fórmula molecular, técnica de medição, nomes químicos, nome comercial, resumo e assunto.

Para Alves e Santos (2013), os metadados de preservação estão relacionados com a conservação e a preservação do objeto descrito. Esse tipo de metadado fornece informações sobre as condições físicas de um recurso, informações sobre as ações tomadas para conservar e preservar as versões físicas e digitais de um recurso. Segundo Weitzel e Mesquita (2015), é importante considerar como proposta de preservação da informação a escolha do formato, tipo de documento, informação sobre autor, título e palavras-chave.

Segundo Lóscio, Burle e Calegari (2017), proveniência significa origem ou fonte de um recurso ou objeto. De acordo com Arakaki (2019, p. 29) os metadados de proveniência "descrevem pessoas, entidades, instituições e atividades envolvidas na produção, influência ou entrega de um dado." Ainda segundo Lóscio, Burle e Calegari (2017), as informações sobre a proveniência dos dados transmitem aos consumidores confiança na integridade e credibilidade dos dados que estão sendo compartilhados. Os metadados de proveniência são fonte, declaração de proveniência, licença, declaração de direitos, titular dos direitos, informações de formato, datas de *copyrights*, data de submissão e de modificação. Além

destes pode-se considerar que os metadados referentes a autorias também transmitem credibilidade ao conjunto de dados.

Os metadados de uso estão relacionados com o nível e tipo de uso dos recursos descritos. As principais tags são: data de criação, controle de uso e usuários e tipo de interação do usuário.

7.2.3 Descrição dos Vocabulários Selecionados

Para estas diretrizes, definiu-se pelo uso integrado de vocabulários amplamente reconhecidos e indicados pelo W3C para descrever e refinar os valores dos metadados, pois conforme Coneglian e Santarem Segundo (2017), a adoção e integração de vocabulários internacionalmente reconhecidos fortalecem a estrutura de descrição semântica dos recursos.

Sendo assim, os vocabulários disponíveis no *Linked Open Vocabularies* (LOV) foram analisados quanto aos conceitos de correspondência das propriedades com os metadados selecionados e decidiu-se pela adoção dos vocabulários Schema.org, DC Terms, SKOS e RDA *Elements Sets*.

O Schema.org fornece uma coleção de vocabulários adotados para estruturar dados para Internet, com vista a melhorar a exibição dos resultados da pesquisa e facilitar a encontrabilidade de páginas na Web. Segundo Laufer (2015) os vocabulários fornecidos pelo Schema.org podem ser utilizados para embutir metadados em páginas Web, as quais são entendidas pelos principais buscadores Google, Microsoft, Yahoo e Yandex. O vocabulário foi fundando pelos mesmos buscadores e seus novos projetos são discutidos no Github por um processo de comunidade aberta para fornecer uma coleção compartilhada de esquemas (SCHEMA.ORG, 2020).

O Schema.org pode ser usado com muitas codificações diferentes, incluindo RDFa, Microdata e JSON-LD. Esse vocabulário abrange entidades, relações entre entidades e ações. Os vocabulários definidos pelo Schema.org seguem uma estrutura hierárquica, sendo que cada um define um tipo. Um *Thing* é o tipo raiz e é relativo a qualquer item genérico. Na página do Schema.org é possível identificar mais de 800 tipos, mais de 1300 propriedades e mais de 300 valores organizados em tipos como *Action, BroadcastService, CreativeWork, Event, Intangible, MedicalEntity, Organizatin, Person, Place e Product*. Cada um desses tipos tem suas próprias especializações, constituindo suas hierarquias. No site do Schema.org é possível consultar o detalhamento de todos os vocabulários.

Nota-se que o Schema.org possui amplas possibilidades de aplicabilidade e contempla propriedades que abarcam as especificidades de uma pesquisa experimental, o foco destas

diretrizes. Dessa forma, as diretrizes estabelecem o Schema.org como vocabulário de descrição dos conjuntos de dados de cadernos de pesquisa em conjunto a termos de metadados de outros vocabulários como o DC Terms, SKOS e RDA *Elemento sets*.

O Dublin Core, mantido pela Dublin Core Metadata Initiative (DCMI), é um padrão de metadados estruturados para descrever recursos informacionais na Web. O DC é "uma das primeiras iniciativas a pensar na definição de vocabulários para a descrição de metadados, tendo como premissas que as descrições tenham independência em relação à sintaxe e tenham uma semântica bem definida" (LAUFER, 2015, p. 92). Foi criado originalmente para promover a descoberta de recursos informacionais na Web, por meio da descrição de identificação mínima, por meio de quinze elementos. O projeto evoluiu e foi desenvolvido um conjunto maior de propriedades e classes, denominados de termos do Dublin Core.

Os termos do Dublin Core (DC Terms), forma abreviada de termos de metadados DCMI, apresentam um conjunto de elementos qualificados que abrange propriedades, esquemas de codificação de vocabulários e de sintaxes e classes que qualificam e refinam os valores dos metadados (DCMI, 2020). Destaca-se que esse padrão considera todos os metadados como opcionais e podem ser repetidos, cabendo à comunidade usuária definir de acordo com a necessidade de cada ambiente informacional quais metadados poderão ou não ser considerados opcionais.

O vocabulário SKOS fornece maneiras de representar sistemas de organização do conhecimento como tesauros, taxonomias, sistemas de controle de autoridade ou qualquer outro tipo de vocabulário controlado e estruturado. Segundo o W3C (2009), o SKOS é uma aplicação RDF que permite que conceitos sejam documentados, ligados e mesclados com outros dados na Web. Assim, objetiva a fácil publicação de vocabulários como *Linked Data*.

De acordo com o W3C (2009), com o SKOS os conceitos podem ser identificados por URIs, com rótulos em uma ou mais línguas e ser documentos com diferentes tipos de notas. Laufer (2015) complementa que os diferentes conceitos podem ser relacionados semanticamente entre si, em hierarquias e redes de associação informacionais, e também agrupados em esquemas conceituais. O SKOS define classes e propriedades que representam os recursos encontrados em um vocabulário controlado. Os principais elementos SKOS, segundo Laufer (2015, p. 96-98), são:

- ***Concept*** – indicado como *skos:concept*, atua como classe que define que um determinado recurso é um conceito.

- ***prefLabel* e *altLabel*** – permitem fazer referência aos conceitos em linguagem natural: *skos:prefLabel* expressa o rótulo preferido e *skos:altLabel* expressa um rótulo alternativo, utilizado, por exemplo, para sinônimos.
- ***Broader* e *narrower*** – o rótulo *skos*:*broader* indica que um conceito é mais abrangente que um outro (um conceito que engloba o outro conceito); o rótulo *skos:narrower* indica um conceito mais específico.
- ***Related*** – o rótulo *skos:related* indica uma relação associativa entre dois conceitos.
- ***Note*** – indica uma observação a respeito do conceito. O rótulo *skos:note* indica uma nota genérica, mas existe a possiblidade de qualificar diferentes tipos de observações, utilizando *skos:scopeNote, skos:historyNote, skos:editorialNote, skos:changeNote*, relativas ao escopo, história, questões editoriais e mudanças efetuadas.
- ***Definition* -** *skos:definition* fornece uma definição do conceito.
- ***ConceptScheme* –** *skos:conceptScheme* atua como classe para indicar que conceitos podem ser criados e usados como entidades independentes.

O autor complementa que o SKOS fornece ainda meios para fazer o mapeamento entre diferentes esquemas conceituais, a partir da relação entre os diversos esquemas.

Os RDA *Element Sets*, [Conjunto de Elementos do RDA] são vocabulários de classes, propriedades e de valores desenvolvidos como parte do código de catalogação RDA. O conjunto de elementos de referência RDA inclui um conjunto de classes que representam as entidades. As principais propriedades do RDA *Element Sets* são: agente, expressão, item, manifestação, *nomen*, lugar, período e obra (RDA, 2019).

7.2.4 Descrição dos Vocabulários Utilizados para Enriquecimento de Dados

Um dos princípios do *Linked Data* é usar URI para nomear coisas, o qual proporciona a identificação fácil de recursos na Web e promove a uniformidade dos recursos. Além disso, destaca-se a importância do reuso de URI como identificadores persistentes dentro dos conjuntos de dados, sempre que possível, para permitir a conexão entre vários *datasets* agregadores e atribuir valores omissos aos dados brutos existentes. Sendo assim, sugere o uso de vocabulários compatíveis com as tecnologias da Web Semântica, tais como:

Quadro 17 - Exemplos de vocabulários utilizados para Enriquecimento de Dados

Vocabulários	Descrição
DBpedia	Descreve recursos em diversos domínios, cujo projeto extrai conteúdo estruturado da Wikipédia. Neste vocabulário é possível identificar agentes como pessoas e instituições.
ORCID *Open Researcher and Contributor* ID	É um código alfanumérico que atua como identificador digital persistente para pesquisadores e outros autores acadêmicos. Apresenta um breve resumo da biografia, país de origem, e-mail do pesquisador, as áreas de interesse e as publicações científicas do pesquisador.
VIAF *Virtual International Authority File*	Atua como identificador persistente digital de pessoas e instituições. O VIAF traz as diferentes formas de apresentação do nome de um autor, em alguns casos as datas de nascimento e morte dos autores, e as diferentes formas de expressão de suas obras.
GeoNames	É um repositório oficial de grafias padrão de nomes geográficos. Este vocabulário pode ser adotado para o estabelecimento de grafias oficiais de nomes estrangeiros, cartografia, sistemas de informação geográfica, inteligência geoespacial e localização espacial de um modo geral.
LCSH *Library of Congress Subject Headings*	Vocabulário controlado para o estabelecimento de assuntos, é mantido pela Biblioteca Nacional dos Estados Unidos e apresenta o controle de assuntos em áreas gerais.
MeSH *Medical Subject Headings*	Apresenta uma estrutura hierárquica de assuntos em Ciências da saúde.
PubChem	É um banco de dados aberto de química no National Institutes of Health (NIH). Contém descrição de moléculas, estruturas químicas, identificadores (InChIkeys e InChI), propriedades químicas e físicas, atividades biológicas, patentes, saúde, segurança, dados de toxicidade, dentre outros.

Fonte: Elaborado pela autora (2020).

Os três primeiro vocabulários – DBpedia, ORCID e VIAF – oferecem identificadores persistentes de agentes para aplicação por meio de metadados de autores principais, colaboradores, co-autores, instituições, agências de fomento e editoras. O GeoNames pode ser aplicado a todos os metadados que remetem à localização geográfica, como lugar de publicação, local da pessoa e instituição pela qual está vinculada, dentre outros. Os vocabulários LCSH, MeSH e PubChem descrevem e controlam autoridades de assuntos.

Esses vocabulários não são os únicos, porém ao selecionar *datasets* para descrição dos vários aspectos de um objeto digital é importante verificar as tecnologias adotadas em sua construção. Os vocabulários exemplificados, além do controle de ambiguidades de nomes de pessoas, instituições e assuntos, adotam tecnologias da Web Semântica, o que permite, por meio *Linked Open Data*, reunir e interligar informações dispersas na Web, contribuindo para o enriquecimento de dados e favorecendo novas descobertas e reutilização de informações.

7.2.5 Definição das Licenças de Uso Utilizadas

Para garantir ao consumidor que o dado possa ser utilizado é indispensável a indicação da licença atribuída a cada conjunto de dados. O propósito de um *Open Notebook Science* é que todos os aspectos de uma pesquisa científica sejam abertos em intervalo de tempo próximo ao real. Essa decisão resguarda a autoria intelectual do pesquisador, ao passo que se algo não está publicado, outras pessoas podem assumir que o cientista não o fez. Entretanto, considerando que existem fatores como receios de lidar com acesso aberto completo e determinados conflitos com propriedade intelectual, apresenta-se os tipos de licença de acordo com o grau de abertura, conforme indicado no quadro 07.

De acordo com as classificações indicadas no quadro 18, cada pesquisador indicará a licença que melhor se adequa ao seu conjunto de dados.

Quadro 18 – Tipos de Licenças Utilizadas

Tipo de Licença	Licença
Domínio público	*Creative Commons* 1.0 Universal (CC0 1.0) *Open Data Commons – Public Domain Dedication and License* (ODC-PDDL)
Atribuição	*Creative Commons – Attribution* (CC-BY) *Open Data Commons – Attribution* (ODC-BY)
Compartilhamento pela mesma licença	*Creative Commons – Attribution – ShareAlike* (CC-BY-SA) *Open Data Commons – Open Database License* (ODC-ODbL)
Com restrições	*Creative Commons – Attribution –NonCommercial* (CC-BY-NC) *Creative Commons – Attribution –NoDerivatives* (CC-BY-ND) *Creative Commons – Attribution –NonCommercial- NoDerivatives* (CC-BY-ND)

Fonte: Elaborada pela autora, a partir do quadro 07.

Apesar de haver possibilidade da indicação de outros tipos de licenças, as de domínio público são especialmente recomendadas para a publicação de dados de pesquisa científica de cadernos de pesquisa, pois facilitam a compilação e o uso massivo da informação. Esta pesquisa não contempla a abordagem para dados comerciais.

7.2.6 Mapeamento das Propriedades de Vocabulários

Após o mapeamento dos metadados (quadro 15), vocabulários (quadro 17) e licenças (quadro 18) que melhor descrevem as especificidades dos dados de pesquisa científica de cadernos de laboratório, realizou-se análise das propriedades dos vocabulários Schema.org, DC Terms e SKOS para identificar correspondências exatas ou aproximadas com os conceitos dos atributos apresentados na primeira coluna dos quadros 19 e 21-23. Os nomes dos atributos

que descrevem os protocolos de pesquisa, os planos de gestão de dados, os *preprints* e *e-prints* foram alinhados para facilitar o entendimento dos usuários.

Quadro 19 - Mapeamento de propriedades de vocabulários - Protocolo de Pesquisa

Planilha de Metadados (Rótulos)	**Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets***	**Tipos de elementos e valores desejados**	
Identificador do registro	schema:identifier	*URI*	Essencial
Data e horário do registro	schema:dateTime	*Date*	Essencial
Autor ▲	schema:author	*Text e URI*	Essencial
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Contribuinte /Colaborador▲	schema:participant	*Text e URI*	
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Instituição▲	schema:sourceOrganization	*Text e URI*	Essencial
Agência de Fomento▲	schema:funder	*Text e URI*	
• Identificador do agente	schema:Identifier	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso variante	skos:altLabel	*String*	
• Campo de atividade	rdaa:P50387	*String*	
• Idioma▲	schema:inLanguage	*String*	
• Informação de contato	schema:email	*Text*	
Título	schema:name	*Text*	Essencial
Subtítulo	schema:alternativeHeadline	*Text*	
Idioma ▲	schema:inLanguage	*String*	Essencial
Formato ▲	dct:format	*String*	Essencial
Tipo▲	dct:type	*String*	Essencial
Número total de páginas	schema:pagination	*Texto*	Essencial
Cobertura espacial▲	schema:location	*URI*	Essencial
Período de execução da pesquisa	schema:startTime	*Date*	Essencial
Público	schema:audience	*Texto*	Essencial
Objetivo pretendido	schema:description	*Texto*	Essencial
Resultados alcançados	schema:result	*Texto*	Essencial
Assunto▲	schema:about	*Text*	Essencial
• Identificador▲	schema:propertyID	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso variante	skos:altLabel	*String*	
• Assunto mais amplo	skos:broader	*String*	
• Assunto mais específico	skos:narrower	*String*	
Descrição	schema:description	*Text*	Essencial

• Reagentes▲	schema:activeIngredient	*Text*	
• InChIKey▲	schema:identifier	*URI*	
• Equipamentos	schema:instrument	*Text*	
• Fórmula molecular▲	schema:identifier	*URI*	
• Peso molecular▲	schema:weight	*String*	
• Técnica de medição	schema:measurementTechique	*Text*	
• Nomes químicos▲	skos:related	*String*	
• Nome comercial▲	skos:related	*String*	
• Data de criação	schema:dateCreated	*Date*	
• Data de modificação	schema:dateModifield	*Date*	
• Período de encerramento da pesquisa	schema:endTime	*Date*	Essencial
• Status da ação	schema:actionStatus	*Text*	Essencial
• Error	schema:error	*Text*	Essencial
Fonte de dados	schema:provider	*Text e URI*	Essencial
Declaração de proveniência	dct:provenance	*Text e URI*	Essencial
Licença de uso	schema:license	*Text e URI*	Essencial
Declaração de direitos	dct:RightsStatement	*Text e URI*	Essencial
Titular dos direitos	dct:rightsHolder	*Text e URI*	
Tamanho do aplicativo	schema:fileSize	*Text*	Essencial
Software necessário	schema:availableOnDevice	*Text*	
Registros de exibição	schema:RegisterAction	*Text*	Essencial
Controle de uso e usuários	schema:userInteractionalCount	*Text*	Essencial
Tipo de interação do usuário	schema:interactionType	*Text*	Essencial

Fonte: Elaborado pela autora (2020).

Na primeira coluna encontram-se os metadados identificados na modelagem (item 7.2.1) e inserção de atributos que descrevem as especificidades dos dados registrados no objeto protocolo de pesquisa, acrescidos de elementos que podem ser enriquecidos com vocabulários externos, como data de nascimento e data de morte de determinado autor, quando for o caso. Os elementos sinalizados com um triângulo (▲) indicam a importância do reuso de informações de *datasets* externos, por meio de URI, sempre que possível, no intuito de evitar ambiguidades, garantir padronização e carregar informações adicionais àquelas requeridas pelos metadados. São exemplos de *tags* de uso de *datasets* externos para enriquecimento de dados dos cadernos de pesquisa:

- Nas *tags* referentes aos autores e colaboradores, instituição e agência de fomento podem ser reutilizadas as informações dos vocabulários ORCID e VIAF, dentre outros, para representar as várias formas do nome dos agentes de acordo com cada idioma adotado.
- Na *tag* de identificação de idioma recomenda o uso das orientações da norma ISO 639-1 que descreve a estrutura, conteúdo, construção e semântica de *tags* para idioma.

- Na *tag* de cobertura espacial recomenda-se o uso do GeoNames, que oferece o identificador do local e traz informações complementares do número de população, fuso horário, informações de latitude e longitude, dentre outras.
- Na *tag* de indicação do formato do arquivo, meio físico ou dimensões do objeto a recomendação é adotar a lista de mídia da Internet, MIME.
- Na *tag* destinada à indicação do tipo do objeto, o DC Terms orienta o uso de vocabulário do tipo DCMI, disponível na página do Dublin Core.
- Nas *tags* de assunto é recomendando o uso do identificador do assunto principal fornecido pelos vocabulários PubChem, MeSH e LSHS.
- A *tag* de descrição é preenchida por meio de *string* textual, mas para seu detalhamento deve ser adotado o vocabulário PubChem, que fornece todos os nomes químicos relacionados, fórmulas moleculares, peso molecular, dentre outras informações. Vale destacar, segundo Bird e Frey (2013), as fórmulas moleculares, embora passíveis de interpretação por máquinas, não são necessariamente únicas e frequentemente são ambíguas. Sendo assim, reforça o uso do identificador único de química para as descrições químicas.

Como já mencionado anteriormente, as *tags* referentes as duas primeiras linhas (identificador do registro e data/horário do registro) e às últimas linhas são preenchidas automaticamente por se tratarem de metadados administrativos e metadados de uso. Destaca que os metadados qualificadores do agente (identificador, ponto de acesso controlado, ponto de acesso variante, campo de atividade, idioma e informação de contato) se aplicam aos autores pessoais, colaboradores, instituição pela qual os autores são vinculados e agência de fomento.

Na segunda coluna encontram-se as propriedades dos vocabulários Schema.org, DC Terms, SKOS e RDA *Elemet Sets*, correspondentes aos metadados da primeira coluna. O Schema.org apresentou o maior número de propriedades com conceitos equivalentes aos metadados mapeados previamente, sendo que as principais classes e propriedades adotadas para a descrição do protocolo de pesquisa são:

Quadro 20 - Classes e propriedades adotadas - vocabulário Schema.Org

Classes/Tipos	Propriedades
Thing	description, identifier, name, alternateName, alternativeHeadline
Person	author, birthDate, deathDate, Identifier, hasOccupation, email, memberOf, participant, inLanguage
Organization	department, funder, sourceOrganization
Action	location, ContentLocation, instrument, result
DataTime	startTime, dateTime
CreativeWork	inLanguage, subjectOf, about, propertyID, sameAs, weight
SoftwareApplication	availableOnDevice, fileSize
Substance	activeIngredient
Product	Weight
PropertyValue	measurementTechnique

Fonte: Elaborado pela autora (2020).

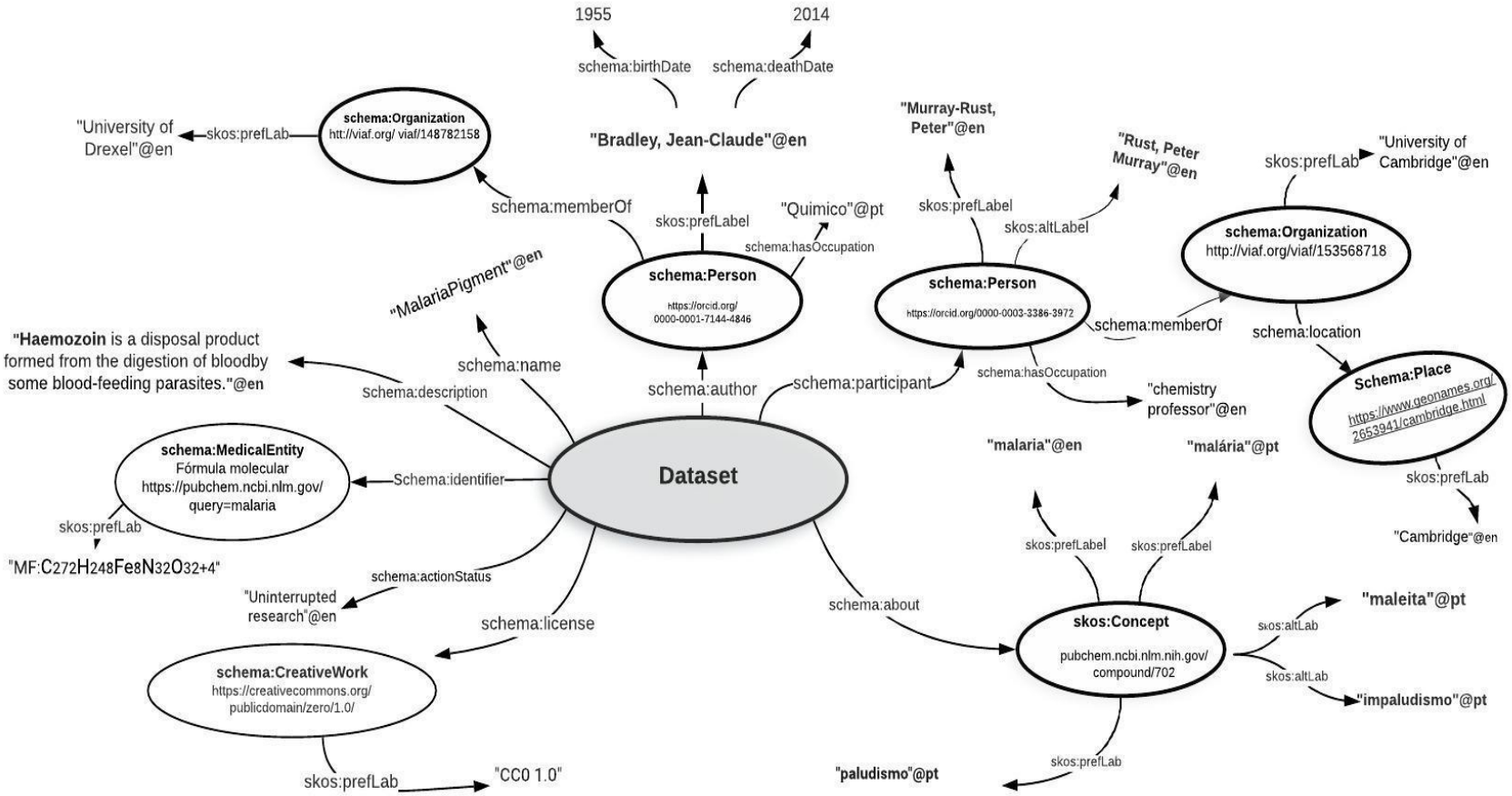
As propriedades são formadas principalmente por uma junção dos padrões Schema.org e DC Terms. Apesar do grande número de propriedades equivalentes, o Schema.org não contempla todas as *tags* de descrição do protocolo de pesquisa, sendo necessário complementar com o DC Terms, modelo genérico que também abrange muitas classes e propriedades necessárias. As principais entidades do DC Terms adotadas foram para indicação do formato, tipo dos dados e informações de proveniências.

Para refinar os metadados, adotou-se as classes e propriedades do vocabulário SKOS, as quais estão declarando rótulos preferidos na indicação de agentes e conceitos dos conjuntos de dados descritos, sendo que a propriedade *skos:prefLabel* indica o ponto de acesso controlado e a *skos:altLabel* indica o ponto de acesso não controlado. O vocabulário SKOS pode sinalizar relações de hierarquia entre os conceitos a partir do uso da propriedade *skos:broader* para indicar o conceito mais amplo e a *skos:narrower* para os conceitos mais específicos. A propriedade *skos:related* sinaliza os conceitos relacionados.

Na terceira coluna são apresentados os tipos de valores desejados aos metadados, sendo URI para indicar o *link* para outro recurso digital, *Date* para informar datas padronizadas, *String* para representar palavras ou textos, por exemplo, “isto é uma *string*”. Ainda na terceira coluna são apresentados os elementos essenciais, considerados como conjunto mínimo para descrição de qualquer tipo de caderno de pesquisa, independente da área de conhecimento pela qual a pesquisa está inserida.

Na figura 28 é possível visualizar a representação de um objeto descrito a partir dos vocabulários indicados, bem como as relações entre as propriedades.

Figura 28 - Representação em grafo do Protocolo de Pesquisa



Fonte: Elaborada pela autora (2020).

A figura 28 ilustra como ocorre a integração dos vocabulários para representação de um objeto no contexto da Web Semântica. O objeto descrito é um Protocolo de Pesquisa que, ao ser publicado em alguma plataforma digital, será representado por URI. A partir desse item principal, a figura demonstra que foram realizadas ligações para outros *datasets* externos que ampliam a descrição e favorecem a recuperação. A recuperação é favorecida na medida em que novas informações não previstas são acrescentadas, como por exemplo, ao utilizar os dados de autoridade do ORCID, que faz referência a Peter Murray-Rust, o usuário terá a oportunidade de recuperar todas as formas do nome do colaborador sem a interferência de ambiguidade.

Nesta representação, é possível observar ainda como são explicitadas as relações entre os elementos descritos por meio do conceito do modelo RDF, o qual atua nessa estrutura como uma base para o processamento dos metadados. Nesse sentido, visualiza a composição de várias triplas do tipo: Recurso, Propriedade e Valor ou Sujeito, Predicado e Objeto. Por exemplo, o protocolo (recurso) foi criado por (propriedade) Jean-Claude Bradley (valor), ou seja, o metadado descritivo *schema:author* ligou o objeto ao autor responsável pelo conjunto de dados, representado pela *string* "Bradley, Jean-Claude"@en. Para realizar essa ligação foi necessário adicionar uma classe para Agente (*schema:Person*), que utilizou dados de autoridade do ORCID para atribuir o valor do recurso ao autor principal "Jean-Claude Bradley". Para indicar a forma preferida do nome do agente na língua inglesa, adotou a propriedade *skos:prefLabel*, enquanto que para informar as datas de nascimento e morte, utilizou as propriedades *schema:birthDate* e *schema:deathDate*, proporcionando mais informações para a busca. Seguindo o exemplo das autorias, tem-se que o objeto está ligado ao colaborador da pesquisa por meio da classe *schema:person* e do metadado *schema:participant*, representado pela *string* "Murray-Rust, Peter@en e também pela outra forma do nome pela *string* "Rust, Peter Murray"@en. Nessa ligação foram adotados dados de autoridade do ORCID para atribuir os valores ao colaborador da pesquisa. Nesse exemplo, nota-se o uso das propriedades *skos:prefLab* para indicar a forma autorizada do nome, representada pela *string* "Murray-Rust, Peter"@en e a propriedade *skos:altLabel* para indicar a forma variante do nome, que por sua vez está representada pela *string* "Rust, Peter Murray"@en, ambas as formas na língua inglesa. Ainda nesse exemplo, foi informado que o colaborador é professor de química por meio do metadado *schema:hasOccupation* e representado pela *string* "chemistry professor"@en.

Na sequência observa que os autores foram ligados, por meio de URI, às instituições pelas quais pertencem. O autor principal Jean-Claude Bradley foi ligado a Universidade de

Dexel. Para realizar essa ligação, adotou-se a classe *schema:Organization* e o metadado *schema:memberOf* para indicar que o autor é membro daquela instituição. Nessa ligação usou os dados de autoridade do VIAF que faz referência à forma preferida da instituição, expressa pela propriedade *skos:prefLab* e representada pela *string* "Univerty of Drexel"@en na língua inglesa. No exemplo do autor colaborador, nota-se também o uso do metadado *schema:memberOf*, que indica que Peter Murray-Rust é membro da Universidade de Cambridge. Para indicar a forma preferida do nome da instituição, denominada de ponto de acesso autoridade, adotou a propriedade *skos:prefLab*, representada pela *string* "University of Cambridge"@en. Destaca-se nesse último exemplo a composição de uma nova tripla que liga a instituição a sua localização geográfica, por meio de um *dataset* externo. Essa ligação foi descrita pelo metadado *schema:location*, representado pela *string* "Cambridge"@en. Para representar o local foi adotada a classe *schema:Place* para atribuir valor ao recurso e foi realizada uma ligação entre o objeto e a fonte externa GeoNames, representada pela propriedade *skos:prefLabel*, que sinaliza o lugar associado ao objeto. Como mencionado anteriormente, esse *dataset* possibilita agregar ao objeto informações de dados geográficos, como nomes de lugares em vários idiomas, elevação e população. O GeoNames está disponível na Web sob a licença *Creative Commons*.

Para atribuir assuntos ao objeto, foram utilizadas informações do vocabulário PubChem por meio da classe *skos:Concept* e representado pelas propriedades *skos:preLabel* para expressar os assuntos preferidos, representados pelas *strings* "malaria"@en e "malária"@pt, nos idiomas da língua inglesa e portuguesa; usou as propriedades *skos:altLab* para indicar os nomes sinônimos "maleita"@pt e "impaludismo"@pt, ambos na língua portuguesa. É importante destacar que o vocabulário PubChem oferece muitas outras informações relativas a pesquisas experimentais realizadas em laboratórios como fórmulas moleculares, os diferentes nomes dos produtos químicos, identificadores universais para nomes químicos, dentre outros.

Uma nova tripla foi formada ao indicar a fórmula molecular adotada na pesquisa experimental. O protocolo, enquanto objeto principal dessa estrutura, está ligado por meio da classe *schema:MedicalEntity*, utilizando o vocabulário PubChem para indicar que a forma adotada da fórmula é "MF:C272H248Fe8N32O32+4". Para realizar essa ligação, usou-se o metadado *schema:identifier* e a propriedade *skos:prefLabel*. Da mesma maneira, adotou-se a classe *schema:CreativeWork* e o metadado *schema:license* para indicar a licença disponibilizada para uso dessa pesquisa.

Outros aspectos relevantes foram a indicação do título (metadado *schema:name*), da descrição da pesquisa (metadado *schema:description*) e do status da pesquisa (metadado *schema:actionStatus*) indicando que a pesquisa é contínua.

Após apresentar a adoção de elementos semânticos estruturados para a publicação de dados de pesquisa científica contidos no protocolo de pesquisa, passa-se para o mapeamento de propriedades de vocabulários do objeto Plano de Gestão de Dados.

Quadro 21 - Mapeamento de propriedades de vocabulários – Plano de Gestão de Dados

Planilha de Metadados	**Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets***	**Tipos de elementos e valores desejados**	
Identificador do registro	schema:identifier	*URI*	Essencial
Data e horário do registro	schema:dateTime	*Date*	Essencial
Autor ▲	schema:author	*Text e URI*	Essencial
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Contribuinte /Colaborador▲	schema:participant	*Text e URI*	
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Nome do gestor dos dados	Schema:name	*Text e URI*	Essencial
Instituição▲	schema:sourceOrganization	*Text e URI*	Essencial
Agência de Fomento▲	schema:funder	*URI*	Essencial
• Identificador do agente▲	schema:Identifier	*URI*	Essencial
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso variante	skos:altLabel	*String*	
• Campo de atividade	rdaa:P50387	*String*	
• Idioma▲	schema:inLanguage	*String*	Essencial
• Informação de contato	schema:email	*Text*	Essencial
Número do subsídio	schema:identifier	*Text*	Essencial
Título	schema:name	*Text*	Essencial
Subtítulo	schema:alternateName	*Text*	
Data da primeira versão	schema:dateCreated	*Date*	Essencial
Data da última atualização	schema:dateModifield	*Date*	Essencial
Período de execução da pesquisa	Schema:startTime	*Date*	Essencial
Formato ▲	dct:format	*String*	Essencial
Tipo▲	dct:type	*String*	Essencial
Idioma ▲	schema:inLanguage	*String*	Essencial
Número total de páginas	Schema:pagination	*String*	Essencial
Cobertura espacial▲	schema:location	*Text e URI*	Essencial
Público	schema:audience	*Text*	Essencial

Objetivo pretendido	schema:description	*Text*	Essencial
Descrição dos tipos de dados pretendidos	schema:description	*Text*	Essencial
Assunto▲	schema:about	*Text*	Essencial
• Identificador do assunto principal▲	schema:propertyID	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso não controlado	skos:altLabel	*String*	
• Assunto mais amplo	skos:broader	*String*	
• Assunto mais específico	skos:narrower	*String*	
• Período de encerramento do experimento	schema:endTime	*Date*	
Fonte	schema:provider	*Text*	Essencial
Declaração de proveniência	dct:provenance	*Text*	Essencial
Licença de uso	schema:license	*Text ou URI*	Essencial
Declaração de direitos	dct:RightsStatement	*Text ou URI*	Essencial
Titular dos direitos	dct:rightsHolder	*Text e URI*	
Políticas éticas	schema:ethicsPolicy	*Text*	
Software necessário	schema:availableOnDevice	*Text*	
Tamanho do aplicativo	schema:fileSize	*Text*	
Controle de uso e usuários	schema:userInteractionalCount	*Text*	Essencial
Tipo de interação do usuário	schema:interactionType	*Text*	Essencial

Fonte: Elaborado pela autora (2020).

As descrições dos elementos que compõem o Plano de Gestão de Dados seguem as mesmas orientações do protocolo de pesquisa, com ressalva para as *tags* de número do subsídio e planejamento de padrões éticos que serão conduzidas nas pesquisas experimentais.

Nota-se, com base em consultas na literatura da área, que um PGD é um documento com menos atributos por se tratar de um planejamento para realização da pesquisa. Sendo assim, não contém os metadados específicos de descrição de experimentos. Nesse caso, esforça-se para indicar o máximo de informações na descrição do resumo, além de informações sobre proveniência e preservação dos dados. Os metadados sobre o número do subsídio (*schema:identifier*) e política de padrões éticas (*schema:ethicsPolicy*), como mencionado, são requeridos em um PGD.

Na terceira coluna encontram-se recomendações de tipos de elementos e valores desejados para descrição de PGD.

A seguir apresenta-se o mapeamento das propriedades de vocabulários referentes ao objeto *Preprint* e *Data Paper*.

Quadro 22 - Mapeamento de propriedades de vocabulários – *Preprint e Data Paper*

Planilha de Metadados	**Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets***	**Tipos de elementos e valores desejados**	
Identificador do registro	schema:identifier	*URI*	Essencial
Data e horário do registro	schema:dateTime	*Date*	Essencial
Autor ▲	schema:author	*Text e URI*	Essencial
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Contribuinte /Colaborador▲	schema:participant	*Text e URI*	
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Instituição▲	schema:sourceOrganization	*URI*	Essencial
Agência de Fomento▲	schema:funder	*URI*	
• Identificador do agente	schema:Identifier	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso variante	skos:altLabel	*String*	
• Campo de atividade	rdaa:P50387	*String*	
• Idioma▲	schema:inLanguage	*String*	
• Informação de contato	schema:email	*String*	
Título	schema:name	*String*	Essencial
• Subtítulo	schema:alternateName	*String*	
Data de produção[26]	schema:dateCreated	*Data*	Essencial
Formato ▲	dct:format	*String*	Essencial
Tipo▲	dct:type	*String*	Essencial
Idioma ▲	schema:inLanguage	*String*	Essencial
Público	schema:audience	*Text*	Essencial
Resumo	schema:abstract	*Text*	Essencial
Assunto▲	schema:about	*Text*	Essencial
• Identificador▲	schema:propertyID schema:sameAs	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso não controlado	skos:altLabel	*String*	
• Assunto mais amplo	skos:broader	*String*	
• Assunto mais específico	skos:narrower	*String*	
• Data de criação	schema:dateCreated	*Date*	
• Data de modificação	schema:dateModifield	*Date*	
• Período de encerramento da pesquisa	schema:endTime	*Date*	Essencial

[26] Data da produção é destinada apenas para recursos não publicados, por exemplo, teses e dissertação.

• Status da ação	schema:actionStatus	*Text*	Essencial
• Error	schema:error	*Text*	Essencial
Palavras-chave	schema:keywords	*Text*	Essencial
Declaração de pré-avaliação	schema:itemReviewed	*Text*	
Fonte de dados	schema:provider	*Text*	Essencial
Declaração de proveniência	dct:provenance	*Text*	Essencial
Licença de uso	schema:license	*Text e URI*	Essencial
Declaração de direitos	dct:RightsStatement	*Text e URI*	Essencial
Titular dos direitos	dct:rightsHolder	*Text*	
Controle de uso e usuários	schema:userInteractionalCount	*Text*	Essencial
Tipo de interação do usuário	schema:interactionType	*Text*	Essencial

Fonte: Elaborado pela autora (2020).

As descrições dos elementos que compõem o *Preprint* e *Data Paper* seguem as mesmas orientações do protocolo de pesquisa quanto a autorias, assunto, proveniências e preservação. Diante das novas demandas apresentas pela ABEC para disponibilização de *preprint* no repositório EmeRI estas diretrizes acrescentaram as seguintes *tags*:

- *Tag* data de produção – para registro da data em que o texto foi postado no repositório de *preprint*.
- *Tag* pré-avaliação por pares - adotou-se o metadado *schema:reviewRating* para informar que a avaliação é dada pela revisão por pares.

O quadro 23 trata do mapeamento das propriedades de vocabulários referentes ao objeto e-*Print*, tido como exemplo o artigo científico.

Quadro 23 - Mapeamento de propriedades de vocabulários – Artigo Científico (*e-Print*)

Planilha de Metadados	**Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets***	**Tipos de elementos e valores desejados**	
Identificador do registro	schema:identifier	*URI*	Essencial
Data e horário do registro	schema:dateTime	*Date*	Essencial
Identificador ▲	schema: identifier	*URI*	Essencial
ISSN	schema:issn	*URI*	Essencial
Autor ▲	schema:author	*Text e URI*	Essencial
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Contribuinte /Colaborador▲	schema:participant	*Text e URI*	
• Data de nascimento	schema:birthDate	*Date*	
• Data de morte	schema:deathDate	*Date*	
• Profissão/Ocupação	schema:hasOccupation	*Text*	
• Instituição vinculada▲	schema:memberOf	*Text e URI*	
• Departamento da Instituição	schema:department	*Text e URI*	
Instituição▲	schema:sourceOrganization	*Text e URI*	
Agência de Fomento▲	schema:funder	*Date*	
• Identificador do agente	schema:Identifier	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	
• Ponto de acesso variante	skos:altLabel	*String*	
• Campo de atividade	rdaa:P50387	*String*	
• Idioma▲	schema:inLanguage	*String*	
• Informação de contato	schema:email	*String*	
Título do artigo▲	schema:headline	*Text*	Essencial
Subtítulo do artigo▲	schema:alternateName	*Text*	
Título do Periódico▲	schema:name	*Text*	Essencial
Designação numérica e/ou alfabética	schema:volumeNumber	*Text*	Essencial
Lugar de publicação▲	schema:location	*Text e URI*	Essencial
Nome da editora▲	schema:publiser	*Text e URI*	Essencial
Data de publicação	schema:datePublished	*Text e URI*	Essencial
• Data de submissão	dct: dateSubmitted	*Date*	
• Data de aceite	dct:dateAccepted	*Date*	
Formato ▲	dct:format	*Text*	Essencial
Tipo▲	dct:type	*Text e URI*	Essencial
Idioma ▲	schema:inLanguage	*Text e URI*	Essencial
Página inicial	schema:pageStart	*String*	Essencial
Página final	schema:pageEnd	*String*	Essencial
Número de páginas	schema:pagination	*String*	Essencial
Público	schema:audience	*Text*	Essencial
Resumo	schema:absctract	*Text*	Essencial
Assunto▲	schema:about	*Text*	Essencial
• Identificador▲	schema:propertyID	*URI*	
• Ponto de acesso controlado	skos:prefLabel	*String*	

• Ponto de acesso não controlado	skos:altLabel	*String*	
• Assunto mais amplo	skos:broader	*String*	
• Assunto mais específico	skos:narrower	*String*	
Fonte de dados	schema:provider	*Text*	Essencial
Declaração de proveniência	dct:provenance	*Text ou URI*	Essencial
Licença de uso	schema:license	*Text ou URI*	Essencial
Declaração de direitos	dct:RightsStatement	*Text ou URI*	Essencial
Titular dos direitos	dct:rightsHolder	*Text e URI*	
Palavras-chave	schema:keywords	*Text*	Essencial
Controle de uso e usuários	schema:userInteractionalCount	*Text*	Essencial
Tipo de interação do usuário	schema:interactionType	*Text*	Essencial

Fonte: Elaborado pela autora (2020).

As diretrizes para os elementos da etapa de autores, assunto e proveniência seguem as mesmas orientações dos demais objetos do ecossistema de cadernos de pesquisa, porém a etapa da descrição física de um *e-Print* assume características de um documento publicado a partir do agente editor. Considerando o artigo científico tomado como exemplo de *e-Print,* destaca-se que é um objeto avaliado por pares e publicado em algum periódico científico, o qual é descrito como parte do artigo. Caso o objeto seja um artigo apresentado em comunicação científica, também foi avaliado por pares e publicado em anais de congresso.

Dito isso, destacam-se as *tag* de metadados destinadas ao artigo científico:

- A *tag* ISSN representada pelo metatado *schema:issn* dará acesso ao periódico no qual o artigo está publicado, preferencialmente que seja indicado por um URI.
- A *tag* título refere-se ao artigo científico, este deve ser descrito preferencialmente por meio da indicação do DOI do próprio texto, por exemplo, o URI https://doi.org/10.19132/1808-5245233.130-156 dará acesso diretamente ao artigo em questão, possibilitando a ligação entre *datasets*, como já mencionado. Em seguida deve-se expressar a forma preferida por meio de *string* textual e propriedade *skos*.
- A *tag* designação numérica e/ou alfabética é a indicação do volume e do número periódico no qual o artigo está publicado, essa informação pode ser informada pelo mesmo DOI e descrito em formato de *string* textual e uma propriedade *skos:prefLabel*.
- A *tag* título do periódico pode ser preenchida pelo URI do ISSN, e assim como os outros elementos anteriormente apresentados, também deve ser expresso em *string* textual para informar a forma preferida do nome.
- A *tag* lugar de publicação é importante que seja informada por meio do identificador geográfico, por meio do vocabulário GeoNames.

- A *tag* nome da editora deve ser representada por um identificador para o nome do editor do periódico, por exemplo, a revista Em Questão é editada pelo Programa de Pós-Graduação em Comunicação da Universidade Federal do Rio Grande do Sul, sendo assim, o VIAF fornece o identificador, via URI, o que favorecerá a ligação do artigo a uma base de autoridade externa.
- Nas *tags* página inicial (*schema:pageStart*), página final (*schema:pageEnd*) e número de páginas (*schema:pagination*) são indicadas em formato de número.

Os demais elementos de um artigo científicos são informados como nos demais objetos.

A seguir apresenta-se o mapeamento de propriedades para relacionamentos entre as diversas partes de um mesmo objeto ou entre diferentes objetos.

7.2.7 Mapeamento de Propriedades para Relacionamento de Agregação

Os objetos digitais descritos assumem várias maneiras para serem publicados, sejam individuais ou em conjunto. Para que sejam publicados a partir de uma abordagem de publicação ampliada, ou seja, a partir da integração de mais de um objeto, surge a necessidade do mapeamento de propriedades que possibilitam interligar um objeto as suas partes complementares, como tabelas, *slides* ou objetos relacionados.

Quadro 24 - Mapeamento de Propriedades de Relacionamento de Agregação

Planilha (rótulo)	Propriedade de os Vocabulários	Descrição
Tem formato	dcterms:hasFormat	Um recurso relacionado que é substancialmente igual ao recurso descrito, mas em outro formato.
É o formato de	dcterms:isFormatOf	Um recurso relacionado pré-existente que é igual ao recurso descrito, mas em outro formato.
Tem parte	dcterms:hasPart	Um recurso relacionado que está incluído físico ou logicamente no recurso descrito.
É parte de	dcterms:isPartOf	Um recurso relacionado no qual o recurso descrito está físico ou logicamente incluído.
Tem versão	dcterms:hasVersion	Um recurso relacionado que é versão, edição ou adaptação do recurso descrito.
É referenciado por	dcterms:isReferenceBy	Um recurso relacionado que fez referência, cita ou, de outra forma, aponta para o recurso descrito.
Relação	dcterms:relation	Um recurso relacionado. A prática recomendada é identificar o recurso relacionado por meio de um URI. Se isso não for possível ou viável, uma *string* em conformidade com um sistema de identificação formal pode ser fornecida.
Sumário	dcterms:tableOfContents	Apresentar uma lista de subunidades do recurso.
Citação	schema:citation	Uma citação ou referência a outro recurso, como outra publicação, páginas da Web, artigo acadêmico etc.

Fonte: Elaborado pela autora (2020).

O uso de vocabulário padronizado permite a agregação de um ou mais recursos da Web ao objeto principal. Uma agregação representa o conjunto de recursos relacionados sobre um objeto real fornecido por provedor de dados (ISAAC, 2014). No contexto dos cadernos de pesquisa, seria considerar o protocolo de pesquisa ou fluxo de trabalho fornecido pelo pesquisador, enquanto provedor de dados, como o objeto principal a ser publicado, e integrá-lo a um artigo científico, por exemplo. É possível agregar o protocolo de pesquisa ao artigo científico por meio da propriedade *dcterms:relation*, seguido de suas descrições e representados por *strings* textuais. Assim como apresentado na figura 28, ao realizar uma agregação adotam-se classes, propriedades de metadados descritivos, propriedades para representação conceitual e dados de autoridade de vocabulários externos, por meio de URIs.

Nessa direção, a agregação pode ocorrer quando o objeto descrito possui uma versão em outro formato de arquivo. Para essa possiblidade, a propriedade adotada é a *dcterms:hasFormat*, direcionando para o segundo formato ou a propriedade *dcterms:isFormatOf* quando a integração ocorre do objeto agregador para o objeto principal.

Segundo Sales (2014), essas agregações, às vezes denominadas de objetos digitais compostos, podem combinar recursos distribuídos com vários tipos de mídia, incluindo textos, imagens, dados, vídeos. Nesse caso, faz-se necessário analisar o conteúdo de cada tipo de mídia para atribuir as propriedades correspondentes.

7.3 ANÁLISE DOS ELEMENTOS SEMÂNTICOS QUANTO AOS PRINCÍPIOS FAIR E MELHORES PRÁTICAS PARA PUBLICAÇÃO DE DADOS NA WEB

Nesta seção serão discutidos o alcance das etapas e os elementos indicados na composição das diretrizes para publicação de cadernos de pesquisa, quanto aos dados a serem encontráveis, acessíveis, interoperáveis e reutilizáveis, a partir da aplicação dos princípios FAIR, tecnologias da Web Semântica e conceitos do *Linked Data*, recomendados pelo W3C.

É importante mencionar que os princípios FAIR, os quais apresentam práticas orientadoras para a boa gestão de dados, vêm sendo requeridos pela comunidade acadêmica, em especial pelas agências de fomento que solicitam em seus planos de gestão que a publicação de dados de pesquisa científica siga os princípios FAIR, conforme descrito na seção 4.4.1. Diante dessa demanda, as diretrizes propostas nesta tese buscaram aplicar os princípios FAIR em conjunto com melhores práticas recomendadas pelo W3C, no que se refere à publicação de dados abertos e ligados ao contexto dos dados de pesquisa científica de cadernos de laboratório.

7.3.1 Encontráveis (*Findable*)

O primeiro princípio FAIR refere-se aos dados e metadados serem encontráveis por pessoas e máquinas, no âmbito da Web. As melhores práticas recomendadas pelo W3C destinadas à encontralidade são o fornecimento de metadados, identificadores persistentes e formatos padronizados legíveis por máquinas, conforme apresentadas no quadro 25.

Quadro 25 - Princípio Findable e Melhores Práticas

FAIR		**Melhores Práticas**	
Encontrável (*Findable*)	F1. (meta) dados são atribuídos a um identificar globalmente exclusivo e persistente	Identificadores de dados	**MP 9** - Usar URIs persistentes de conjunto de dados.
			MP 10 - Usar URIs persistentes como identificadores dentro de conjuntos de dados.
			MP 11 - Atribuir URIs a versões e séries de conjuntos de dados.
	F2. os dados precisam ser descritos com metadados	Metadados	**MP 1** – Fornecer metadados para humanos e máquinas.
	F3. os metadados incluem clara e explicitamente o identificador dos dados que descreve		**MP 2 –** Fornecer metadados descritivos.
			MP 3 – Fornecer metadados estruturados.
	F4. (meta) dados são registrados ou indexados em um recurso pesquisável	Formato de dados	**MP 12** – Usar formatos padronizados legíveis por máquina.

Fonte: Silva, Santarem Segundo e Silva (2018, p. 91).

Nas diretrizes estabelecidas para a estruturação dos cadernos de pesquisa foram adotados identificadores globais, únicos e persistentes para descrever nomes de objetos, pessoas, instituições, lugares e assuntos, por meio da definição de *tags* de metadados.

As melhores práticas do *Linked Data* esclarecem que os indicadores assumem muitas formas e são usados em todos os sistemas de informação. Vale destacar que a descoberta, o uso e a citação de dados na Web dependem fundamentalmente do uso de URIs HTTP que podem ser consultados na internet. Sendo assim, as MP 9, 10 e 11 sinalizam para o uso de URI persistente para conjunto de dados e URI como identificadores de conjuntos de dados, além de atribuir URI para as versões e séries de conjuntos de dados. Nesse sentido, os quadros 19, 21-23 apresentam *tags* direcionadas para o uso de identificadores ou URIs, como identificador de objeto digital (DOI) para representar artigos científicos, identificadores de pessoa por meio de vocabulários como o ORCID e VIAF, indicadores para cobertura espacial por meio do vocabulário GeoNames e indicadores de assunto por meio dos vocabulários LCSH, MeSH e PubChem. Este último oferece o identificador internacional de química (InChIKey) e descreve nomes químicos, fórmulas moleculares, compostos químicos, dentre outros utilizados em muitas pesquisas experimentais nas diversas áreas do conhecimento.

De acordo com as melhores práticas do W3C, o uso correto de identificadores oferece aos dados os benefícios de reuso, ligação, descoberta e interoperabilidade.

Quanto aos dados serem descritos utilizando metadados ricos, foram selecionados aqueles que descrevem e enriquecem os dados a partir de *datasets* externos. O princípio encontrável (*findable*) recomenda que um conjunto de dados deva ser descrito por metadados ricos o suficiente para que, uma vez indexados em um mecanismo de busca, possam ser encontrados mesmo sem o seu identificador persistente. As MP 1, 2 e 3 recomendam fornecer metadados descritivos, estruturais e administrativos. Atendendo as essas recomendações, as diretrizes foram construídas a partir de mapeamentos de metadados e esquemas de metadados para descrever os dados de pesquisa relacionados à pesquisa experimental, não pretendeu ser demasiadamente exaustivo e sim descrever as informações consideradas suficientes para que o pesquisador possa analisar e optar pela reprodução ou repetibilidade dos dados.

Para fornecer metadados descritivos, os pesquisadores, enquanto provedores e editores de dados, têm a importante contribuição de fornecer seus conjuntos de dados com informações completas e corretas para facilitar a compreensão, uso e reuso dos dados.

No contexto de enriquecimento de dados Henning *et al.* (2019) reforçam que os metadados enriquecidos devem incorporar propriedades sobre os dados, viabilizando a execução de diversas tarefas rotineiras, como atualização das formas do nome dos agentes,

verificação licenciamento, limpeza e manutenção de conteúdo de forma automática, com o desenvolvimento de tarefas que exigem muita atenção dos pesquisadores. Nesse sentido, entende-se que o enriquecimento de dados se dá pela inclusão de *datasets* externos, como os mencionados na seção 7.2.4 (Descrição dos vocabulários utilizados para enriquecimento de dados).

Para a publicação de dados de pesquisa científica, no contexto semântico, é necessário oferecer metadados interpretáveis pelos usuários humanos e máquinas. Na ocasião de implementação das diretrizes faz-se necessário oferecer metadados para leitura humana, conforme Lóscio, Burle e Calegari (2017) recomendam, fornecendo metadados como parte de uma página da Web HTML e como um arquivo de texto separado. Enquanto que para a interpretação de máquinas, os metadados podem ser fornecidos em um formato de serialização Turtle e JSON, ou podem ser incorporados na página HTML (HTML-RDFA ou JSON-LD) e reutilizar padrões existentes e vocabulários populares, como por exemplo o padrão de metadados Dublin Core e Schema.org. Dessa forma, optou-se pelo uso integrado dos padrões de metadados Schema.org e Dublin Core para possibilitar a descrição detalhada dos valores dos metadados. O Schema.org foi selecionado para descrever os objetos digitais em torno dos cadernos de laboratório por possuir o maior número de propriedades que correspondem com as pesquisas experimentais e o Dublin Core, que possui múltiplas propriedades para descrever informações como as proveniências, formatos e tipos dos dados. Além destes, adotou-se nas diretrizes o vocabulário SKOS para refinar os valores dos metadados, como assuntos gerais, específicos e relacionados.

De acordo com Lóscio, Burle e Calegari (2017), os metadados possibilitam os benefícios do reuso, compreensão, descoberta e processabilidade, enquanto que os princípios FAIR destacam a importância do uso de metadados em todas os seus princípios.

Em F3 recomenda-se que os metadados incluam explicitamente os identificadores dos dados que descreve. Isto porque os metadados e o conjunto de dados são geralmente apresentados em arquivos separados. Assim, a orientação é explicitar a associação entre metadados e conjunto de dados por meio de identificador persistente, conforme apresentados nos princípios F1 e F2.

A prática recomendada em F4 é que os metadados sejam registrados e indexados em mecanismos de busca. Nesse sentido, os princípios anteriores (F1, F2 e F3) orientaram quanto aos metadados para a descrição detalhada dos objetos digitais. No entanto, para executar a tarefa de busca e recuperação é necessária a implementação desses elementos em uma plataforma digital. Na interpretação de Silva, Santarem Segundo e Silva (2018), a MP 12

colabora com o princípio F4, pois recomenda o uso de formatos padronizados e legíveis por máquinas ao publicar dados na Web. Na explicação dos autores, na medida em que os dados se tornam mais onipresentes e os conjuntos de dados se tornam maiores e mais complexos, o processamento por computadores torna-se indispensável. Sendo assim, espera-se que as máquinas possam ler e processar os dados publicados na Web e os humanos possam usar ferramentas disponíveis no domínio relevante para trabalhar com eles. Entre os formatos indicados incluem, mas não limita, a sintaxe de serialização CSV, XML, HDF5, JDON e RDF como RDF/XML, JSON ou Turtle (LÓSCIO; BURLE; CALEGARI, 2017).

O quadro 26 apresenta alguns exemplos de usos de identificador único e persistente de pessoas, instituição, local e assunto (F1), dados enriquecidos com *datasets* externos (F2 e F3) e esquemas de metadados populares e amplamente utilizados (F2) contemplados nas diretrizes semânticas propostas para a estruturação de dados de pesquisa de cadernos de laboratório.

Quadro 26 - Exemplos de alguns elementos FAIR adotados nas diretrizes

Metadados	**Esquemas de Metadados**	**Valores enriquecidos com identificadores**
Autor	schema:author	https://orcid.org/0000-0003-0729-3339
Instituição	schema:sourceOrganization	http://viaf.org/viaf/156757416
Agência de Fomento	schema:funder	http://viaf.org/viaf/147398725
Cobertura espacial	schema:location	https://www.geonames.org/3448439/sao-paulo.html
Fórmula molecular	schema:identifier	https://pubchem.ncbi.nlm.nih.gov/#query=C2H6O
Composto químico	skos:related	https://pubchem.ncbi.nlm.nih.gov/compound/702
InChIKey identificador universal de química	schema:identifier	LFQSCWFLJHTTHZ-UHFFFAOYSA-N
Licença de uso	schema:license	http://creativecommons.org/publicdomain/zero/1.0/

Fonte: Elaborado pela autora (2020).

Além das melhores práticas mencionadas por Silva, Santarem Segundo e Silva (2018, p. 91), os autores Rautenberg, Souza, Dall'Agnol e Michelon (2018) recomendam a MP 24 para usar padrões Web como base das APIs e MP 35 para citar o publicador original como forma de proporcionar o benefício da descoberta. A MP 24 poderá ser contemplada no momento da implantação dessas diretrizes. A MP 35 foi adotada nas diretrizes por meio da *tag schema:provider* para que o pesquisador possa indicar a fonte dos dados, a qual é um dos principais indicadores de confiabilidade de dados.

7.3.2 Acessível *(Accessible)*

A proposta deste princípio é que os dados sejam acessados por outros pesquisadores. Segundo Wilkinson *et al*. (2016), a acessibilidade dos dados está relacionada ao uso de protocolos de comunicação padronizados, abertos e gratuitos, que ofereçam autenticação e acesso aos metadados mesmo quando não estiver mais disponível.

Quadro 27 - Princípio *Accessible* e Melhores Práticas

FAIR		**Melhores Práticas**	
Acessível (*Accessible*)	A1. (meta) dados são recuperáveis pelo seu identificador usando um protocolo de comunicação padronizado	Acesso a dados	**MP 23** – Disponibilizar dados por meio de uma API.
	A1.1 o protocolo é aberto, gratuito e universalmente implementável		**MP 24** – Usar padrões da Web como base de API.
	A1.2 o protocolo permite um procedimento de autenticação e autorização, quando necessário		*Não foram identificadas diretrizes específicas que atendem este requisito.
	A2. os metadados são acessíveis, mesmo quando os dados não estão mais disponíveis		**MP 22** – Fornecer uma explicação para dados que não estão disponíveis (*MP Adaptada para a faceta A2*).

Fonte: Silva, Santarem Segundo e Silva (2018, p. 94).

Ao se planejar a implementação dessas diretrizes, a recomendação é possibilitar o acesso ao conjunto completo de dados de uma determinada pesquisa. Segundo Silva, Santarem Segundo e Silva (2018) esse conjunto pode ser constituído por dados de pesquisa primários e também por resultado de uma publicação ampliada composta por dados de pesquisa, tese ou dissertação, artigos de revistas e de comunicações de eventos. Para o fácil acesso a esses conjuntos de dados, a MP 17 recomenda que a infraestrutura da Web deva ser implantada de modo a permitir o acesso em massa de um conjunto de dados completo com apenas um pedido, evitando inconsistência no acesso individual de dados ao longo de muitas recuperações, bem como permitir o fornecimento de subconjuntos de dados (MP 18), caso os consumidores não precisem do conjunto completo. Para a implementação de um conjunto de dados que contenham vários arquivos é necessário tornar os dados acessíveis para *download* por meio de um URI (LÓSCIO; BURLE; CALEGARI, 2017).

Ainda segundo Lóscio, Burle e Calegari (2017), por padrão, a Web oferece acesso usando métodos de protocolo de transferência de hipertexto (HTTP) para *download* simples, em massa, de um arquivo. Ainda que os dados estejam distribuídos em vários URIs, estes podem ser organizados em um modelo de contêiner, através do protocolo de transferência de arquivos, para facilitar o acesso em massa aos dados. Os dados distribuídos em vários arquivos podem ser recuperados através de uma interface de programação de aplicativos

(API), método de recuperação mais sofisticado. Uma API geralmente é a abordagem mais flexível para servir subconjuntos de dados, pois permite personalizar aqueles que são transferidos, tornando os subconjuntos disponíveis muito mais prováveis de fornecer os dados necessários (MP 18). De acordo com a MP 23, uma API geralmente é a abordagem mais flexível e oferece capacidade de processamento para os consumidores de dados. A MP 20 indica o uso de API para o fornecimento de dados em tempo real e a MP 23 indica a API para ativar o uso de dados em tempo real. No caso de dados de pesquisa científica, o tempo dependerá do tipo de pesquisa que está sendo realizada. O acesso a dados de uma publicação ampliada poderá não ocorrer em tempo real em função dos trâmites das defesas públicas, avaliações por pares e publicações em repositórios (SILVA, SANTAREM SEGUNDO, SILVA, 2018).

A MP 24 reforça que as APIs criadas com base nos padrões da Web aproveitam pontos fortes, como por exemplo, usar verbos HTTP como métodos e URIs que mapeiam diretamente para recursos individuais e ajudam a evitar um acoplamento rígido entre solicitações e respostas (LÓSCIO; BURLE; CALEGARI, 2017).

O princípio A2 orienta que os metadados devam ficar acessíveis, mesmo quando os dados não estão disponíveis, enquanto que a W3C orienta, por meio da MP 22, fornecer explicações para dados que não estão disponíveis, informando sobre como podem ser acessados e quem pode acessá-los. Nesse caso, as metodologias FAIR e MP podem se complementar disponibilizando metadados, mesmo quando os dados não estiverem disponíveis, e ainda assim oferecer mensagens explicativas. Uma possível abordagem de sua implementação é a publicação de um documento HTML que forneça tais explicações legíveis para a indisponibilidade de dados. As mensagens podem ser apresentadas a partir de códigos de status HTTP, como por exemplo as mensagens 303 (ver outro), 410 (permanentemente removido), 503 (serviços indisponíveis), dentre outros (SILVA, SANTAREM SEGUNDO, SILVA, 2018).

Além dessas, Lóscio, Burle e Calegari (2017) recomendam a MP 32 para atender o beneficio do acesso aos dados. A MP 32 foi contemplada nessas diretrizes por meio da *tag schema:license*, além da indicação do quadro 18 com os tipos de licenças.

7.3.3 Interoperável *(Interoperable)*

A interoperabilidade refere-se à capacidade de um sistema se comunicar facilmente com outro. Para isso, faz-se necessária a adoção de vocabulários de representação e

metadados interligados. Silva, Santarem Segundo e Silva (2018) reforçam a necessidade de identificadores persistentes, metadados e o reuso de vocabulários padronizados. Além disso, Rautenberg, Souza, Dall'Agnol e Michelon (2018) expõem a importância do uso de API para disponibilizar dados (MP 23), usar padrões Web como base das APIs (MP 24) e evitar alterações significativas na API (MP 26). Dessa forma, a seguir, serão descritos os princípios aplicados na elaboração das diretrizes para estruturação de dados de cadernos de pesquisa.

Quadro 28 - Princípio *Interoperable* e Melhores Práticas

<table>
<tr><th colspan="2">FAIR</th><th colspan="2">Melhores Práticas</th></tr>
<tr><td rowspan="3">Interoperável
(Interoperable)</td><td>I1. (meta) dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento</td><td rowspan="2">Vocabulários de dados</td><td>MP 15 - Reutilizar vocabulários, preferencialmente padronizados para codificar dados e metadados.</td></tr>
<tr><td>I2. (meta) utilizam vocabulários e/ou ontologias que seguem os princípios de FAIR</td><td>MP 16 – Escolher o nível de formalização correto.</td></tr>
<tr><td>I3. (meta) incluem referências qualificadas para outros (meta) dados</td><td>Acesso a dados</td><td>MP 22 – Fornecer uma explicação para dados que não estão disponíveis (MP adaptada).</td></tr>
</table>

Fonte: Silva, Santarem Segundo e Silva (2018, p. 97).

Uma linguagem para ser formal necessita do fornecimento de metadados legíveis por humanos e máquinas (MP 1 e 2), uso de indicadores persistentes (MP 9 e 10) e o reuso de vocabulários padronizados (MP 15 e 16). Silva, Santarem Segundo e Silva (2018) indicam que os vocabulários proporcionam a integração dos dados ao definirem conceitos e relacionamentos entre as entidades usadas para descrever e representar uma área de interesse. A recomendação de Lóscio, Burle e Calegari (2017) é a de fazer uso de vocabulários amplamente utilizados e padronizados para codificar dados e metadados.

Com o intuito de estruturar os cadernos de pesquisa, buscou-se na iniciativa *Linked Open Vocabularies* (LOV) os vocabulários que descrevessem seus propósitos, como a descrição de autoridades com refinamento de atributos que indicassem as datas, a profissão, o campo de atividade, os meios de comunicação com os agentes e a possiblidade de vincular os pesquisadores às instituições e departamento dos quais fazem parte. Além disso, buscou-se vocabulários que descrevessem as especificidades de uma pesquisa experimental, como nome de compostos químicos, propriedades químicas, procedimentos realizados durante a pesquisa, períodos de realização, status da pesquisa e uma maneira de informar se a pesquisa foi bem sucedida ou não; buscou também vocabulários que possibilitassem a descrição de proveniências dos dados. Após o detalhamento dos atributos que descrevem as especificidades dos cadernos de pesquisa, optou-se pelos vocabulários Schema.org, DC Terms

e SKOS, além dos vocabulários de valores como GeoNames, VIAF, ORCID, dentre outros exemplificados em F1 a F3. Como modelo de representação, a prática recomendada é o uso do RDF e suas serializações, também mencionada em F2. Este último deverá ser adotado no momento de implementação das diretrizes. Ademais, o uso de vocabulários amplamente reconhecidos favorece os benefícios interoperabilidade, processabilidade, compreensão, confiabilidade e o reuso dos dados.

Silva, Santarem Segundo e Silva (2018) atendem o princípio I3, acerca de referências qualificadas entre metadados, e recomendam as orientações da MP 22 para o uso de uma explicação contextual sobre a situação dos dados e de seus metadados. A implementação pode ocorrer por meio da publicação de um documento HTML, com explicações legíveis para a situação dos dados.

7.3.4 Reutilizável *(Re-Usable)*

Este princípio visa aumentar as possibilidades de reuso de um conjunto de dados por diferentes consumidores de dados. Trata-se de um princípio especialmente importante para o contexto dos cadernos de pesquisa, pois reflete a aplicação dos princípios anteriores (encontrável, acessível e interoperável) e ao objetivo final de todo o processo de estruturação, que se refere ao reuso dos dados por diferentes pesquisadores em novas pesquisas. Foram contemplados nas diretrizes propostas para estruturação de cadernos de pesquisa os seguintes elementos reutilizáveis.

Quadro 29 - Princípio *Re-Usable* e Melhores Práticas

FAIR		**Melhores Práticas**	
Reutilizável (Re-usable)	R1. (meta) dados são ricamente descritos com uma pluralidade de atributos precisos e relevantes	Metadados	**MP 1**– Fornecer metadados.
			MP 2 – Metadados descritivos.
			MP 3 – Metadados estruturais.
	R1.1. (meta) dados são liberados com uma licença de uso de dados clara e acessível	Licença de dados	**MP4** – Fornecer informações de licença de dados.
	R1.2. (meta) dados estão associados com a proveniência detalhada	Proveniência de dados	**MP5** – Fornecer informações completas sobre as origens dos dados e quaisquer alterações feitas.
	R1.3. (meta) dados cumprem os padrões comunitários relevantes para o domínio		*Não identificadas diretrizes específicas para este requisito.

Fonte: Silva, Santarem Segundo e Silva (2018, p. 98).

Considerando que o princípio R1 requer que os metadados sejam descritos com pluralidade de atributos precisos e relevantes, as diretrizes mapearam os metadados para os dados de pesquisa experimental com base nos fundamentos bibliográficos, que buscam

representar os principais aspectos de um objeto digital, contemplando metadados de proveniência, de descrição das especificidades dos cadernos de pesquisa, administrativos e de uso. Além daqueles para serem descritos com valores textuais, o uso de identificadores persistentes foi amplamente recomendado nessas diretrizes para fins de enriquecimento de dados. Segundo Wilkinson et al. (2016), esse detalhamento de atributos permite ao pesquisador avaliar a possibilidade do seu reuso, bem como adequações as suas necessidades.

Conforme o princípio R1.1 os metadados devem ser disponibilizados com licenças de uso acessível. Nesse sentido, as MP 4 e MP 34 destacam a importância de fornecer uma licença para evitar limitações de reuso e formalizar legalmente a disponibilização de dados para reutilização do trabalho de outra pessoa. Nestas diretrizes, a *tag schema:license* foi disponibilizada para indicar uma das licenças descritas no quadro 18. No entanto, a prática recomendada é que o pesquisador responsável pelos dados e metadados indique uma licença de domínio público para ampliar o uso das informações.

Segundo o princípio R1.2, os metadados devem estar associados à sua proveniência. No contexto dos cadernos de pesquisa, o conceito de proveniência visa assegurar a integridade e a credibilidade dos dados. A MP 5 recomenda que o provedor de dados forneça um nível apropriado de detalhes sobre a origem de seus dados. Nestas diretrizes foram mapeadas as *tags* para fonte dos dados (*schema:provider*), para indicar a origem dos dados; declaração de proveniência (*dct:provenance*) para explicitar quaisquer alterações na propriedade e custódia de um objeto; licença (*schema:license*) para permitir fazer o uso em relação ao objeto; declaração de direitos (*dct:RigtsStatement*) para informar sobre os direitos intelectuais e conceder a permissão oficial visando fazer uso do objeto; titular dos direitos (*dct:rightsHolder*) para indicar o nome do agente que possui ou gerencia os direitos do objeto; data de criação (*schema:dateCreated*) e modificação dos dados (*schema:dateModifield*) para informar a data original da primeira versão e das modificações feitas nos dados. Portanto, ao declarar com detalhes as técnicas e procedimentos adotados nos experimentos, também que sejam informações de proveniência e parâmetros de avaliação e qualidade dos dados.

Além das melhores práticas correlacionadas no quadro 29 desta tese, Lóscio, Burle e Calegari (2017) recomendam as MP 12 a 14, referentes aos formatos de dados para obter os benefícios da processabilidade e do reuso de dados. A MP 12 aconselha o uso de formatos padronizados e legíveis por máquinas para permitir a interoperabilidade e reusos futuros, conforme exemplificados em F2 desta seção.

A MP 13 orienta disponibilizar informações sobre parâmetro de localidade, para evitar dificuldades de compreensão dos dados que se modificam de um idioma para outro, como por

exemplo, datas e moedas; o recomendável é usar estruturas e valores de dados localmente neutros ou fornecer metadados sobre a localidade usada pelos valores de dados, como por exemplo *dc:language* do padrão Dublin Core. O valor pode ser conectado ao catálogo de códigos para representação de nomes de idiomas da *International Organization for Standardization* (ISO) 639, publicada pela *Library of Congress* como <dc:language xsi:type="dcterms:ISO639-2">en</dc:language>. Nestas diretrizes, como a opção foi pelo Schema.org o exemplo seria *schema:inLanguage.*

A MP 14 recomenda sempre que possível a disponibilidade de dados em múltiplos formatos, quando mais de um se adequa ao seu uso pretendido. Ao implementar as diretrizes propostas nesta tese, recomenda-se a conversão dos dados para os formatos RDF/XML, JSON e Turtle, os quais atendem à recomendação das MP 12 de uso de formato legível por máquina, MP 14 formatos em múltiplos formatos, MP 15 vocabulários conhecidos e padronizados.

Lóscio, Burle e Calegari (2017) recomendam ainda as categorias qualidade, indicadores de versão, preservação, *feedback* e republicação de dados para melhorar o reuso ao publicar dados na Web. Essas categorias não são indicadas nos Princípios FAIR.

8 CONSIDERAÇÕES FINAIS

O desafio de concluir esta pesquisa demanda um movimento de retomá-la em seu início e, então, relatar a síntese dos resultados alcançados. A proposta desta tese teve como ponto de partida a constatação de que o ecossistema da pesquisa cientifica vem passando por diversas transformações que interferem no modo como a informação é gerada, disponibilizada e compartilhada. Essa transformação está ocorrendo, em âmbito mundial, por uma gama de fatores, dentre os quais, podem-se destacar os avanços na tecnologia, exigência por produção acadêmica, pressões de financiamento e cultura colaborativa entre pesquisadores. Junto a esses fatores observa-se um cenário em que as informações de qualidade produzidas no domínio acadêmico e científico estão disponibilizadas de forma desestruturadas, isoladas e de difícil acesso no ambiente Web, isto em plena era da *e-Science*. Ainda nessa linha de reflexão, observa-se, nos últimos anos, que a exigência por produção científica se estendeu da publicação de resultados de pesquisas para publicação dos dados primários, sobretudo às áreas que envolvem estudos experimentais, com a ideia de acelerar novas descobertas e prover economia de recursos advindos de financiamento público. Essa exigência vem ocorrendo em vários países e inclusive no Brasil, principalmente pelas agências de fomento.

Nessa perspectiva, observou-se a existência de repositórios para o armazenado de dados de pesquisa realizados em laboratórios. Mesmo assim, muitas instituições de ensino e pesquisa estão iniciando seus próprios repositórios de dados de pesquisa como meio para o armazenamento, preservação e acesso aos dados. Dessa forma, vislumbrou-se nas tecnologias da Web Semântica o caminho para a estruturação e publicação de tais dados.

Diante de tais constatações e do conhecimento do potencial das tecnologias da Web Semântica em reaproveitar dados existentes na Web e proporcionar significados à informação, surgiu o seguinte questionamento: ***Como estruturar dados de pesquisa científica anotadas em cadernos de pesquisa para publicá-los, em formato aberto e semântico, que atenda as demandas da e-Science?*** Esse questionamento se desmembrou em: Quais elementos conceituais e práticos presentes nas dimensões da *e-Science* precisam ser observados para publicação de dados de pesquisa científica? Quais recomendações e tecnologias semânticas podem ser adotadas para publicar conjuntos de dados de cadernos de pesquisa em plataformas de acesso aberto? São questionamentos que levaram à definição do objetivo geral: Propor um conjunto de diretrizes semânticas para estruturar e publicar dados de pesquisa científica registradas em cadernos abertos de pesquisa, visando a melhorias na recuperação e no compartilhamento de dados em plataformas de acesso aberto.

Para responder aos questionamentos e alcançar o objetivo geral desta tese, foram definidos cinco objetivos específicos, os quais serão citados e comentados na sequência.

O primeiro objetivo consistiu em **analisar, com base na literatura, os elementos conceituais e práticos presentes nas dimensões da *e-Science***. Após a descrição das dimensões em um formato de pirâmide – dados de pesquisa científica, tecnologia e colaboração científica em rede – foi possível atender ao objetivo da pesquisa.

Nesse sentido, identificou que a dimensão 'dados de pesquisa científica', que se situa na base mais baixa e maior da pirâmide da infraestrutura da *e-Science*, é composta pela enorme quantidade de dados decorrentes das novas gerações de tecnologias e pesquisas científicas. Nesta dimensão, os elementos identificados foram as ações gerenciais, as boas práticas, os padrões e tecnologias destinados para a gestão dos dados.

A segunda dimensão analisada foi a 'tecnológica', na qual foi identificado que os elementos necessários para a construção de projetos que se propõem a se inserir no contexto do movimento *e-Science* - e que se encontram dentro do escopo desta tese - são as linguagens e padrões da Web Semântica e boas práticas do *Linked Data* para representação e ligação de uma variedade de anotações. A tecnologia mais recomendada foi a estrutura de descrição de recursos, RDF, a qual codifica o significado de etiquetas por meio de incontáveis triplas. Esse sistema padrão oferece um grau de flexibilidade muito mais alto ao tipo de metadados que pode ser armazenado em comparação com os bancos de dados relacionais tradicionais.

A terceira dimensão analisada - 'colaborações em redes' - requer um conjunto de tecnologias distribuídas que permitem que pessoas geograficamente distantes interajam e colaborem para a construção de um conhecimento comum. Os principais elementos de apoio à colaboração científica em rede são as plataformas digitais e aplicativos de TICs.

O capítulo 3 apontou, nas três referidas dimensões, que os registros de dados de pesquisa ocorrem mediante a descrição de metadados relevantes, por meio de informações sobre a proveniência, o conteúdo e as condições em que os dados foram produzidos. A literatura atribui grande importância ao fornecimento de metadados, pois a publicação de dados de pesquisa sem metadados estruturados e definidamente anotados ocasiona a omissão de interpretação e é possível que os dados sejam perdidos. Diante da importância da atribuição de metadados no processo de organização e publicação de dados, os autores citados em todo o capítulo enfatizam que a devida escolha dos esquemas de metadados é necessária para a integração entre instrumentos, experimentos e laboratórios.

O segundo objetivo específico se propôs a **apresentar as características e especificidades dos dados de pesquisa científica anotados em cadernos de laboratório**, o

qual foi atendido por meio do estudo dos capítulos 4 e 5, que, ao eleger o protagonismo dos cadernos de pesquisa no processo de publicação de dados de pesquisa científica, fez-se necessária a compreensão de como são constituídos, quais dados são publicados neste instrumento, quais suas características e especificidades, para então propor a estruturação de tais dados. Sendo assim, essa compreensão abarcou os estudos distribuídos no capítulo 4, a partir de uma abordagem geral sobre o estudo dos dados de pesquisa científica como um todo; e os estudos do capítulo 5 sobre as especificidades dos cadernos.

A partir do estudo desses capítulos, compreendeu-se que o principal aspecto que diferencia os dados registrados em cadernos de pesquisa dos demais tipos de dados de pesquisa científica está em suas características e tipologias. A literatura mostrou que os dados gerados em laboratórios são classificados como dados experimentais, os quais são resultados de procedimentos realizados em condições controladas com a finalidade de provar o estabelecimento de hipótese sobre determinado fenômeno. Foi possível constatar que na classificação ‘dados experimentais’ são incluídas informações sobre procedimentos realizados nos experimentos e resultados de estudos de laboratório, como medições de reações químicas. Os cadernos de pesquisa podem conter números, imagens, fluxos de vídeo ou áudio, informações sobre a versão do software, algoritmos, equações, animações ou simulações, conforme apresentados no quadro 12. Os dados experimentais podem ser combinados ou gerados novos.

O terceiro objetivo específico consistiu em **analisar, a partir de iniciativas existentes, as práticas favoráveis e as desfavoráveis na publicação de dados de pesquisa científica anotados em cadernos abertos de pesquisa**, o qual foi contemplado ainda no capítulo 5. Para tanto, realizou-se consultas em repositórios de dados de pesquisa e em *blogs* para conhecer, na prática, as especificidades dos dados experimentais e as práticas favoráveis e as desfavoráveis, a partir das orientações apresentadas no capítulo 4. Para isso, os elementos metadados, formatos, padrões e grau de estrutura dos dados foram consultados nas plataformas dos cadernos *UsefulChem, LabScribbles* e *Openlabnotebooks.*

O *UsefulChem,* projeto descontinuado, foi um bom exemplo de caderno de pesquisa, no sentido de oferecer metadados descritivos essenciais ao contexto de uma pesquisa experimental. No entanto, não foi possível identificar nenhum tipo de vocabulário adotado. A plataforma em que os dados eram postados era um *blog*. Os dados publicados na plataforma *Blogger* não são estruturados, pois não há um modelo de dados identificável e legível por máquinas, impossibilitando que eles sejam extraídos, transformados e processados. Dessa forma, a contribuição desse caderno refere-se ao apoio na definição de metadados descritivos.

Os dados de pesquisa que os cadernos *LabScribbles* e *Openlabnotebooks* publicam seguem uma estrutura baseada em vocabulários e formatos legíveis por usuários humanos e máquinas, o que facilita o uso e compartilhamento de dados. Porém, em alguns casos, ao acessar o documento normalmente em formato de texto, foram identificados gráficos em formato de foto (JPEG e PNG), o que poderia ser apresentada também em planilhas (CSV, ODS, XLS, XLSX), pois os dados quando disponibilizados em tabelas facilitam o seu processamento e sua recuperação por ferramentas computacionais. Além dessas fragilidades, observou-se que os cadernos descrevem minimamente os objetos, portanto entende-se que os diferentes tipos de objetos poderiam contemplar a descrição mais detalhada dos metadados para oferecer aos usuários melhores chances de encontrá-los. Dessa maneira, não foi possível identificar metadados referentes à proveniência de dados, tais como fonte de dados, declaração de proveniência e declaração de direitos os quais transmitem credibilidade ao conjunto de dados.

O quarto objetivo específico **identificou os conceitos e tecnologias da Web Semântica e *Linked Data* para publicar conjuntos de dados de cadernos de pesquisa em plataformas de acesso aberto** – apesar dos capítulos 4 e 5 apresentarem recomendações para estruturação e publicação de dados abertos por meio dos princípios FAIR e das melhores práticas recomendadas pelo W3C, não descreveu as tecnologias da Web Semântica e *Linked Data*. Sendo assim, o capítulo 6 se dedicou em analisar e identificar as tecnologias que contemplam cada prática recomendada pelo Consórcio W3C. Entre as tecnologias, incluem RDF, URI, serializações RDF, ontologias, padrões de metadados e o modelo conceitual IFLA LRM. Essas tecnologias, aos serem aplicadas ao contexto dos princípios FAIR e melhores práticas, proporcionam aos cientistas e consumidores de dados os benefícios de reuso, compreensão, interligação, descoberta, confiança, acesso, interoperabilidade e processabilidade dos dados.

Após atender aos quatro objetivos específicos, o capítulo 7 apresentou um **conjunto de diretrizes semânticas para estruturação e publicação de dados de pesquisa científica de cadernos de laboratório**, contemplando o quinto objetivo específico desta tese, que consistiu em identificar etapas e elementos semânticos para compor tais diretrizes.

As diretrizes semânticas, neste estudo, referem-se ao conjunto de orientações elaboradas com base nas tecnologias da Web Semântica e conceitos do *Linked Data* para estruturação e publicação de cadernos abertos de pesquisa. Essas diretrizes constituem a proposta final desta tese e acredita-se ter obtido êxito, conforme apresentadas a seguir.

Para definir as diretrizes fez-se necessário conhecer o ecossistema de dados de pesquisa em torno da pesquisa experimental, o qual permitiu identificar que esse ecossistema é formado não só por objetos como também por pessoas e organizações. No contexto dos cadernos de pesquisa destaca-se que os principais atores são os pesquisadores, na qualidade de autor, colaborador ou orientador de uma pesquisa, e as instituições das quais os atores fazem parte, bem como os laboratórios em que os experimentos são realizados. No momento oportuno esses agentes com o apoio de um bibliotecário de dados poderão ser importantes na implementação das diretrizes. Enquanto que os objetos que compõem o ecossistema são o caderno de pesquisa, formado por protocolos de pesquisa, ou seja, os principais objetos de descrição de uma pesquisa, pois contém o registro dos procedimentos realizados, equipamentos, instrumentos químicos, fórmulas utilizadas, produtos e medidas químicas, além de todas as informações referentes a pesquisa. Indentificou-se que o plano de gestão de dados é um objeto importante no contexto dos cadernos de pesquisa, pois descrevem todas as etapas de ciclo de vida dos dados, os metadados que preveem a preservação dos dados e a documentação do processo de pesquisa. Além destes, incluem os *preprints*, *data papers* e os *e-prints*. Os *preprints* podem oferecer interpretações recentes dos autores da pesquisa, pois são publicados rapidamente antes da avaliação pelos pares. Os *data papers* também oferecem informações atualizadas, porém sem inferências dos autores. Os *e-prints* podem ser livros, capítulos de livros, artigos científicos de periódicos ou comunicação científica, dentre outros documentos já publicados. Essa integração de dados agrega valor ao produto final e possibilita melhorias na qualidade da recuperação.

Em seguida, passou-se ao estudo das etapas e de elementos para compor as diretrizes. A definição seguiu as orientações do modelo fluxo organizacional para publicação de dados, no que se refere à estruturação de dados, de Santarem Segundo (2018), com atenção às orientações dos modelos de ciclo de vida de dados do DataONE e DCC*Curation*, dos princípios FAIR e das melhores práticas do W3C. As etapas são:

Etapa 1 – a modelagem dos conjuntos de dados buscou descrever formalmente os vários aspectos dos dados de pesquisa em torno do caderno de laboratório e explicitar os relacionamentos entre eles. A proposta do modelo conceitual IFLA LRM foi adotada nessa modelagem, a qual foi relevante na definição dos metadados para descrever os conjuntos de dados, a partir das necessidades de informação dos usuários e na definição de relacionamentos entre os elementos presentes nesses cadernos, favorecendo o processo de mapeamento de propriedades.

Etapa 2 – o mapeamento de dados proporcionou a identificação dos metadados que visam descrever e individualizar o protocolo, enquanto principal objeto do caderno de pesquisa, além do mapeamento dos metadados dos objetos plano de gestão de dados, *preprint* e artigo científico, como exemplo de *e-print*, conforme apresentado no quadro 15, que servirão para uma futura publicação ampliada desses objetos em conjunto com os cadernos de pesquisa. Após mapear os metadados, foi apresentada a descrição de cada atributo e suas recomendações de uso.

Etapa 3 – descreveu os vocabulários selecionados na análise de correspondências com os metadados identificados no mapeamento de dados. Os vocabulários descritos foram: Schema.org, o qual oferece milhares de propriedades para estruturar dados na Internet com possibilidades de aplicação em diferentes áreas do conhecimento; o DC Terms, que apresenta uma gama de propriedades para descrição de objetos digitais na Web, no entanto, por apresentar características mais gerais, não contemplou muitas propriedades para estruturação de cadernos de pesquisa. Ainda assim, os metadados referentes à proveniência de dados serão representados pelo dcterms; o SKOS foi usado para representar esquemas de controle de autoridades como pessoas e assuntos. Nessas diretrizes, as propriedades SKOS foram amplamente adotadas para refinar metadados e indicar formas autorizadas e variantes dos nomes.

Etapa 4 – na etapa da descrição dos vocabulários utilizados para enriquecimento de dados, buscou-se encorajar os publicadores de dados a fazerem uso, sempre que possível, de URIs como forma de identificar coisas, para que haja a ligação com outros recursos, evitando a ambiguidade de nomes e atribuindo valores omissos aos dados.

Etapa 5 – na etapa da definição das licenças de uso foi apresentado um quadro com os nomes das licenças para que o pesquisador proprietário dos dados indique, por meio do metadado *schema:license*, aquela que melhor corresponde a sua pesquisa. A apresentação da licença é uma forma de garantir ao consumidor que o dado possa utilizar determinado conjunto de dados. Relembra aqui que a proposta de Jean-Claude Bradley para o *Open Notebook Science* é de disponibilizar os dados abertamente sem restrição de uso o mais próximo do tempo real. Dessa forma, reforça a importância de disponibilizar conjunto de dados sob licença do tipo domínio público para facilitar a compilação e o uso da informação.

Etapa 6 – o mapeamento das propriedades de vocabulários deu-se pela análise de correspondência entre os conceitos dos atributos e as propriedades dos vocabulários selecionados para estruturação e posterior publicação dos dados de objetos digitais. Os principais vocabulários que apresentaram maior aderência foram o Schema.org, DC Terms e

SKOS, como já mencionados. São três vocabulários amplamente reconhecidos e recomendados pelas melhores práticas do W3C. As classes e propriedades desses vocabulários foram trabalhadas de forma integrada com a intenção de fortalecer a estrutura da descrição semântica dos recursos. Além dos quadros 19, 21 a 23, a figura 28, exemplifica a representação de um objeto descrito (protocolo de pesquisa) a partir dos vocabulários indicados, bem como as relações entre as classes e propriedades. Nessa representação é possível observar como são explicitadas as relações entre os elementos descritos por meio do conceito do modelo RDF, que atua nessa estrutura como uma base para o processamento dos metadados. Vale destacar que a figura 28 mostra como ocorrem as ligações com *datasets* externos, por meio de URI, ampliando a descrição com o reaproveitamento de dados existentes na Web, enriquecendo as informações dos dados principais e favorecendo a descoberta das informações.

Etapa 7 – o mapeamento de propriedades para relacionamentos foi realizado para possibilitar a integração entre partes de um objeto ou entre vários objetos, no momento da implementação das diretrizes. Nesse sentido, mapeou as propriedades a partir dos vocabulários Schema.org e DC Terms.

Após o mapeamento do ecossistema de dados de pesquisa do caderno de laboratório e da descrição das etapas e elementos de estruturação e publicação de dados de pesquisa científica de cadernos de laboratório, realizou-se análise dos elementos quanto ao alcance dos dados de serem encontráveis, acessíveis, interoperáveis e reutilizáveis, a partir da aplicação dos princípios FAIR, das tecnologias da Web Semântica e dos conceitos do *Linked Data*.

O resultado obtido foi que o fornecimento de metadados administrativos, descritivos, de proveniência, de preservação e de uso pode garantir que dada pesquisa de cadernos de laboratórios sejam **encontráveis**. Isto com a indicação de implementação a partir de vocabulários padronizados e que permitam a leitura por pessoas e máquinas. Ademais, há demasiada recomendação da atribuição de URIs para indicação de nomes e enriquecimento de dados.

Somando-se às recomendações de uso de APIs, as quais poderão favorecer a **acessibilidade** aos dados. Além desses elementos, a recomendação de uso de uma explicação contextual sobre a situação dos dados e seus metadados pode proporcionar a **interoperabilidade** dos dados. Destaca que, juntando as recomendações anteriores à indicação de licença de uso, preferencialmente, que seja de domínio público, bem como o detalhamento de metadados associados à proveniência dos dados, facilita o **reuso** dos dados de pesquisa dos cadernos de laboratórios publicados a partir de tais diretrizes. Vale salientar

que nem todas as recomendações relacionadas aos princípios da acessibilidade estão contempladas nessas diretrizes de estruturação, no entanto fica a recomendação para a ocasião de implementação.

Após percorrer pelas etapas da pesquisa registra-se que as principais contribuições dessas diretrizes estão na reunião de elementos e tecnologias que refletem a realidade de pesquisas laboratoriais e a descrição de experimentos com uma pluralidade de atributos ricos, precisos e relevantes, os quais poderão proporcionar benefícios à comunidade cientifica com dados organizados, padronizados e disponíveis para o reuso. A aplicação devida dessas diretrizes, no que se refere à estruturação de dados, garantem que eles sejam encontráveis, acessíveis, interoperáveis e reutilizáveis com segurança, à medida que os dados serão devidamente citados por meio de metadados de proveniência. Considera-se que este estudo esteja alinhado aos fundamentos da Ciência da Informação, a começar pela, visto que relaciona tecnologias e conceitos de organização da informação advindos de duas áreas diferentes que se complementam - computação e biblioteconomia - e resultar em um conjunto de diretrizes semânticas que visam à organização e representação da informação, desde a sua origem até o seu reuso.

Nessa perspectiva, entende-se que a implementação das diretrizes propostas nesta tese seja o inicio de um novo ciclo de organização da informação, no âmbito da Web, pois ao aplicar as diretrizes novas questões relacionadas às especificidades de laboratórios poderão surgir, sendo necessárias adaptações e aperfeiçoamentos. Nesse aspecto, esclarece que o mapeamento de dados e análise de correspondência de entidades de vocabulários e metadados foram realizados com base em cadernos de pesquisa experimental de natureza genérica, podendo atuar como um teste inicial e abrir possibilidades para a estruturação de cadernos temáticos. Isto porque os cadernos temáticos podem assumir especificidades não contempladas nessas diretrizes. Sendo assim, novas aplicações podem ser planejadas e implementadas conforme necessidades da comunidade usuária de tais diretrizes.

Por fim, espera-se que o conjunto de diretrizes semânticas colabore para o processo de descrição e recuperação da informação, apoie as novas práticas científicas, contribua para a modernização da Ciência da Informação e para o avanço da Ciência.

8.1 DESAFIOS ENFRENTADOS

O maior desafio no desenvolvimento desta pesquisa foi trabalhar com um tema que vem sendo discutido do ponto de vista teórico, mas que ainda é compreendido com receio e preocupação por parte dos pesquisadores de laboratórios e das próprias instituições em aceitar

que seus dados possam ser disponibilizados com segurança, do ponto de vista da propriedade intelectual, e que novos resultados possam ser gerados a partir destes. Observou-se que muitos pesquisadores, mesmo com financiamento público, possuem o sentimento de posse dos dados e acreditam que uma pesquisa deva ser divulgada apenas em suas configurações finais, ou seja, o resultado da pesquisa via artigo científico. Essa constatação ocorreu a partir da primeira proposta de estudo de realizar a pesquisa em um laboratório de determinada área do conhecimento, dentro de uma instituição de ensino superior para, então, mapear e caracterizar todos os tipos de dados produzidos dentro de um laboratório de pesquisa e, na sequência, propor um modelo de estruturação e publicação de tais dados. No entanto, apesar das várias tentativas, a proposta inicial foi declinada e reconduzida para estudos baseados em experiências relatadas na literatura e estudo dos próprios cadernos disponíveis na Internet. Diante dessa constatação, considera-se que ações de conscientização e incentivo sejam necessárias para o avanço da proposta de publicação de dados de pesquisa no contexto dos cadernos de laboratórios, em âmbito nacional.

8.2 SUGESTÕES PARA TRABALHOS FUTUROS

São sugeridas as seguintes propostas de pesquisa para estudos futuros:

1 – Implementar as diretrizes semânticas construídas nesta tese e sinalizar os pontos positivos e negativos encontrados na proposta.

2 – Levantamento de cadernos de pesquisa por áreas temáticas, com a intenção de conhecer as especificidades das áreas, bem como identificar os tipos de dados produzidos por estes laboratórios para, então, mapear atributos e vocabulários capazes de descrever e individualizar cada conjunto de dados de pesquisa.

3– Diante da importância da integração de dados de diferentes objetos digitais em torno dos cadernos de pesquisa para fins de ampliar a oferta de informação de qualidade aos pesquisadores, enquanto consumidores de dados, com foco na aceleração de novas descobertas, sugere-se o estudo com maior profundidade de um modelo de publicação ampliada para cadernos de pesquisa.

4 – Considerando que alguns dos argumentos para publicar dados de pesquisa de cadernos de laboratório em formato aberto se basearam em colaborar para novas pesquisas e economia de recursos públicos, surge a proposta de identificar e reunir as políticas de abertura de dados de pesquisa científica desenvolvidas no Brasil e no exterior, e medir a efetividade

dessas politicas, bem como comprovar o impacto social e econômico da abertura desses dados.

REFERÊNCIAS

ABDO, Alexandre Hannud. Ciência aberta, da ciência para todos à ciência com todos. **Liinc em Revista**, Rio de Janeiro, v. 10, n.2, p.460-471, nov. 2014. Disponível: http://revista.ibict.br/liinc/article/view/3592/3071. Acesso em: 22 jun. 2020.

ALBAGLI, Sarita; APPEL, André Luiz; MACIEL, Maria Lúcia. E-Science, ciência aberta e o regime de informação em ciência e tecnologia. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v.7, n.1, jan./jun. 2014.

ALBAGLI, Sarita; MACIEL, Maria Lucia; ABDO, Alexandre Hannud (Orgs). **Ciência aberta, questões abertas**. Brasília, DF: IBICT; Rio de Janeiro: UNIRIO, 2015. 312 p.

ALLARD, Suzie. DataONE: protecting the future of environmental and ecological data. **Proceedings of the American Society for Information Science and Technology**, v. 46, n.1, p 1-5, 2010.

ALMEIDA, Cleibson Aparecido de *et al*. Melhoria na qualidade de dados com a aplicação de "data cleaning" na base de dados de acidentes aeronáuticos da aviação civil brasileira. **AtoZ**: novas práticas em informação e conhecimento, n. 5, v.2, p.72-79, jul./dez. 2016.

ALVES, Rachel Cristina Vesu; SANTOS, Plácida Leopoldina V. A. C. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013. 196 p.

ALVES, Rachel Cristina Vesu. **Metadados como elementos do processo de catalogação**. Orientadora: Plácida Leopoldina Ventura Amorim da Costa Santos. 2010. 132 f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho, 2010.

ANDRONICO, Giuseppe *et al*. E-Infrastructures for e-science: a global view. **J Grid Computing**, v. 9, p.155-184, 2011. DOI 10.1007/s10723-011-9187-y

APPEL, Andre Luiz. **A e-Science e as atuais práticas de pesquisa científica**. Orientadora: Maria Lúcia Maciel. 2014. 88 f. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Informação, Instituto Brasileiro de Informação e Tecnologia. Universidade Federal do Rio de Janeiro, 2014.

APPEL, Andre Luiz; MACIEL, Maria Lucia; ALBAGLI, Sarita. A e-Science e as novas práticas de produção colaborativa de conhecimento científico. **Revista Internacional de Ciencia y Sociedad**, v.3, n. 1, 2016. Disponível em: http://repositorio.ibict.br/bitstream/123456789/928/1/470-1274-1-PB.pdf. Acesso em: 18 out. 2019.

ARAKAKI, Felipe Augusto. **Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV**. Orientadora: Plácida Leopoldina Ventura Amorim da Costa Santos. 2019. 139 f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho, 2019.

ARAÚJO, Alberto Ávila. **O que é ciência da informação**. Belo Horizonte: KMA, 2018. 131 p.

ARAÚJO JÚNIOR, Rogério Henrique. **Precisão no processo de busca e recuperação da informação**. Brasília, DF: Thesaurus, 2007. 175 p.

ASSOCIATION OF RESEARCH LIBRARY. Joint Task Force on Library Support forcience. **Agenda for Developing E-Science in Research Libraries**. 2007. 26 p. Disponível em: https://www.arl.org/wp-content/uploads/2007/12/escience-report-final-2007.pdf. Acesso em: 12 abr. 2018.

BACON, Dave. **Pseudo open notebook science?** Publicado no blog "The Quantum Pontiff" em 26 jun. 2008. Acervo pessoal cedido David Bradley, via e-mail.

BALL, Alex. **Review of data management lifecycle models**. Bath: University of Bath, 2012. 14 p. Disponível em: https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf. Acesso em: 02 fev. 2020.

BELL, G. Foreword. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. **The fourth paradigma**: data intensive scientific discovery. Washington: Microsoft Research Redmond, 2009.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, 17 may 2001.

BERNERS-LEE, T. **Linked Data**. 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 13 jun. 2018.

BERNERS-LEE, T. Is your Linked Open Data 5 star? *In*: BERNERS-LEE, T. **Linked Data**. Cambridge: W3C, 2010. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em 29 mar. 2020.

BERNERS-LEE, Tim. **What do HTTP URIs identify?** 2002. Disponível em: https://www.w3.org/DesignIssues/HTTP-URI.html. Acesso em: 20 mar. 2020.

BIRD, Colin L; FREY, Jeremy G. Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences. **Chem. Soc. Rev**. n.42, 2013. DOI: 10.1039/C3CS60050E.

BIRD, Colin L.; WILLOUGHBY, C.; FREY, Jeremy G. Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences. **Chem Soc. Rev**., n. 42, v.20, out. 2013. Doi: 10.1039 / c3cs60122f

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data: the story so far. **International Journal on Semantic Web and Information Systems**, v.5, n.3, p. 1-22, 2009. Disponível em: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. Acesso em: 09 out. 2017.

BORKO, Harold. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

BORGMAN, C. Research data: who will share what, with whom, when, and why? In: CHINA--NORTH AMERICAN LIBRARY CONFERENCE, 5., 2010, Beijing. Disponível em: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_161.pdf . Acesso em: 13

jun. 2018.

BRADLEY, Jean-Claude . **Open notebook science claims and logos**. Blog Useful-Chem. Publicado em: 24 fev. 2009. Disponível em: http://usefulchem.blogspot.com.br/2009/02/open-notebook-science-claims-and-logos.html. Acesso em: 02 nov. 2019. Acervo pessoal cedido David Bradley, via e-mail.

BRADLEY, Jean-Claude. **Blog UsefulChem Experiments**. Experimento 013. 31 maio 2006. Disponível em: http://usefulchem-experiments1.blogspot.com/2006/05/exp-013.html. Acesso em: 31 out. 2019.

BRADLEY, Jean-Claude. **Open notebook science, reproducibility and exclusion**. 2009. Disponível em: usefulchem.blogspot.com/2009/02/open-notebook-science-reproducibility.html.

BRADLEY, Jean-Claude. **Open Notebook Science**. Experimento 045. 27 de março de 2007.

BRADLEY, Jean-Claude. **Opening up and sharing** [Internet] Chemistry World; 2013, abril 18. Disponível em http://www.rsc.org/chemistryworld/2013/04/open-science-chemistry-sharinginformation

BRADLEY, Jean-Claude. **Peer review and Science 2.0**: blogs, wikis and social networking sites. Guest lecture for Peer review culture in scholarly publication and Grantmaking" course at Drexel University em 15 mar. 2010. Disponível em: <http://usefulchem.blogspot.com.es/2010/03/peer-review-and-science20-talk.html Acesso em: 05 nov. 2014.

BRADLEY, Jean-Claude. The impacto of open notebook Science. **Information Today**, [S.l], v. 27, n.8, p. 50-51, set. 2010. Entrevista realizada por Richard Poyder. Disponível em: http://www.infotoday.com/it/sep10/Poynder.shtml#top. Acesso em: 02 set. 2019.

BRADLEY, Jean-Claude; LANG, Andrew S. I. D.; KOCH, Steve; NEYLON, Cameron. Colaboration using open notebook science in academia. In: **Collaborative computational technologies for biomedical research**. John Wiley & Sons, 2011. p. 425-452. Disponível em: http://media.wiley.com/product_data/excerpt/36/04706380/0470638036-1.pdf. Acesso em: 02 set. 2019.

BRADLEY, Jean-Claude. **Jean-Claude Bradley Drexel University and blogmaster of usefulchem.blogspot.com**. Entrevista para David Bradley. Blog. Postagem em: jan. 2006. Disponível em: https://www.reactivereports.com/51/51_0.html . Acesso em:

BRADLEY, Jean-Claude. **Open Notebook Science using blogs and wikis**. Slides. Palestra ministrada em: 27 mar 2007.

BRADLEY, Jean-Claude. **Open notebook Science**. Blog. Publicado em: 26 set. 2006. Disponível em: http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html. Acesso em: 25 maio 2019.

BRADLEY, Jean-Claude; OWENS, Kevin; WILLIAMS, Antony. Chemistry crowdsourcing and open notebook Science. **Nature Precedings.** 9 jan. 2008. DOI: 10.1038/npre.2008.1505.1

BREITMAN, K. K.; CASANOVA, M. A.; TRUSZKOWSKI, W. **Semantic web**: concepts, technologies and applications. Cham: Springer, 2007. Disponível em: http://dx.doi.org/10.1007/978-1-84628-710-7. Acesso em: 02 mar. 2020.

BURTON, Adrian; TRELOAR, Andrew. Designing for discovery and re-use: the "ANDS Data Sharing Verbs" approach to service decomposition. **International Journal of Digital Curation**, n. 3, v. 4, p.44–56, dez., 2009. ISSN: 1746-8256. DOI: 10.2218/ijdc.v4i3.124.

CASTELLS, Manuel. **A sociedade em rede**. 6. ed. São Paulo: Paz e Terra, 2006. 698 p.

CESAR JUNIOR, Roberto Marcondes. Apresentação à edição brasileira. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo: Oficina de Textos, 2011, p.7-8.

CHEN, P. **Modelagem de dados**: a abordagem entidade-relacionamento para projeto lógico. Tradução de Cecília Camargo Bartalotti. São Paulo: McGraw Hill, 1990.

CLINIO, Anne. **Novos cadernos de laboratório e novas culturas epistêmicas**: entre a política do experimento e o experimento da política. Orientadora: Sarita Albagli. 2016. 240 f. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Instituto Brasileiro de Informação e Tecnologia. Universidade Federal do Rio de Janeiro, 2016.

CLINIO, Anne. Por que open notebook science? Uma aproximação às ideias de Jean-Claude Bradley. In: ALBAGLI, Sarita; MACIEL, Maria Lucia; ABDO, Alexandre Hannud (Orgs). **Ciência aberta, questões abertas**. Brasília, DF: IBICT; Rio de Janeiro: UNIRIO, 2015. p. 253-286.

CLINIO, Anne; ALBAGLI, Sarita. Cadernos abertos de laboratório e publicações líquidas: novas tecnologias literárias para uma Ciência Aberta. **Rev Eletron Comun Inov Saúde**, nov. 11. 2017. ISSN 1981-6278.

CONEGLIAN, Caio Saraiva *et al*. Repositório GPNTI-CRIS: um ambiente para publicação científica ampliada. *In*: ENCUENTRO DE LA ASSOCIACIÓN E INVESTIGACIÓN EM CIENCIA DE LA INFORMACIÓN DE IBEROAMÉRICA Y EL CARIBE, 11. 2018. **Anais eletrônicos** ... Medellín, Colombia, 2018.

CONEGLIAN, Caio Saraiva; SANTAREM SEGUNDO, José Eduardo. Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais. **Encontros Bibli: revista eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 22, n. 48, p. 88-99, jan./abr. 2017. ISSN 1518-2924. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n48p88. Acesso em: 09 out. 2017.

COSTA, Maíra Murrieta. **Diretrizes para uma política de gestão de dados científicos no Brasil**. Orientador: Murilo Bastos da Cunha. 2017. 288 f. Tese (Doutorado em Ciência da Informação) — Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília, Brasília, DF, 2017.

CRESWELL, John W. **Projeto de pesquisa**: métodos qualitativos, quantitativos e misto. 3. ed. Porto Alegre: Artmed; Bookman, 2010.

CROWSTON, Kevin; QIN, Jian. A Capability maturity model for scientific data management: evidence from the titerature. **Asist&t**, v. 45, n.1, jan. 2012. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/meet.2011.14504801036.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de biblioteconomia e arquivologia**. Brasília, DF: Brinquet de Lemos/Livros, 2008. 451 p.

CURTY, Renata Gonçalves. O paradigma da publicação de dados e suas diferentes abordagens. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília, SP. **Anais eletrônicos** ... Marília, SP. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/468/820. Acesso em: 02 nov. 2019.

CURTY, Renata Gonçalves. **Beyond "Data Thrifting"**: na investigation of factors influencing research data reuse in the social sciences. Orientador: Jian Qin. 2015. 266f. Tese (Doutorado em Filosofia) – School of Information Studies [Escola de Estudos da Informação], Universidade de Syracuse (NY), 2015. Disponível em: https://surface.syr.edu/etd/266. Acesso em: 17 set. 2017.

DIGITAL CURATION CENTRE. **DCC Curation lifecycle model**. 2020. Disponível em: http://www.dcc.ac.uk/resources/curation-lifecycle-model. Acesso em: 05 jan. 2020.

DEMO, Pedro. **Metodologia científica em ciências sociais**. 3. ed. rev. e ampl. São Paulo: Atlas, 2012.

DREXEL UNIVERSITY. College of Arts and Sciences. **Jean-Claude Bradley, PhD, Departamento de Química**. (Mensagem memorial). 2014. Disponível em: https://drexel.edu/coas/news-events/news/2014/May/mourning-jean-claude-bradley-department-of-chemistry/. Acesso: em 24 jul. 2019.

DUTTON, W. **Collaborative network organizations**: new technical, managerial and social infrastructures to capture the value of distributed intelligence. Oxford: Oxford Internet Institute, 17 nov. 2008. DPSN Working Paper Series, n.5.

FARIA-CAMPOS, Alessandra C. *et al*. Protocol data management in biology laboratories: proposal for the development of na information management system. **Brazilian Journal of Information Studies: Research Trends**. V. 10, n. 1, 2020, p. 173-189. Disponível em: http://revistas.marilia.unesp.br/index.php/bjis/article/view/9174/6287. Acesso em: 13 jul. 2020.

FERREIRA, Valdinéia Barreto. **E-Science e políticas públicas**: ciência, tecnologia e inovação no Brasil. Salvador: EDFBA, 2018. 255 p.

FORCE11. The Future of Research Communications and e-Scholarship. **Guiding principles for findable, accessible, interoperable and reusable data publishing version B1.0**. 2014. Texto digital. Disponível em: < https://www.force11.org/fairprinciples>. Acesso em: 13 jun. 2018.

FOSKETT, D. J. Informática. *In*: GOMES, Hagar Espanha. (Org.). **Ciência da informação ou informática?**. Rio de Janeiro: Calunga, 1980. p. 1-51. (Ciência da Informação).

FOSTER. **Open Science**. 2018. Disponível em: <https://www.fosteropenscience.eu/foster>. Acesso em: 09 jan. 2020.

FUSCO, Elvis. **Aplicação dos FRBR na modelagem de catálogos bibliográficos digitais**. São Paulo: Cultura Acadêmica, 2011.

GALVÃO, Taís Freire; PEREIRA, Mauricio Gomes. Revisões sistemáticas da literatura: passos para sua elaboração. **Epidemiol. Serv. Saúde** [on-line]. 2014, vol.23, n.1, pp.183-184. ISSN 1679-4974.

GARCIA, Joana Coeli Ribeiro; TARGINO, Maria das Graças. Open peer review sob a ótica de editores das revistas brasileiras da Ciência da Informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília, SP. **Anais eletrônicos do XVIII Enancib**. Marília, SP. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/19. Acesso: 13 set. 2019.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2016.

GRAY, Jim. Jim Gray on eScience: a transformed scientific method. *In*: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. **The fourth paradigma**: data intensive scientific discovery. Washington: Microsoft Research Redmond, 2009. Palestra. Editada por Tony Hey, Stewart Tansley, Kristin Tolle.

GRAY, Jim; SZALAY, Alex. **eScience**: a transformed scientific method. Palestra proferida na Computer Science and Telecommunications Board (CSTB). 2007. 59 slides. Disponível em: https://sites.nationalacademies.org/CSTB/index.htm. Acesso em: 09 nov. 2018.

GRAY, J. eScience: a transformed scientific method. Transcrição de palestra ministrada por Jim Gray no Conselho Nacional de Pesquisa (EUA), 11 Jan. 2007. In: HEY, T.; TANSLEY, S.; TOLLE, K (Ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond: Microsoft Research, 2009.

GREEN, Ann; MACDONALD, Stuart; RICE, Robin. **Policy-making for research data in Repositories**: a guide. May 2009. Disponível em: https://www.coar-repositories.org/files/guide.pdf. Acesso em: 01 out. 2019.

GOLD, Anna. Cyberinfrastructure, data, and libraries. **D-Lib Magazine**, Massachutts, v.13, n.9/10, sept./ oct. 2007. Parte 1: Cyberinfrastructure primer forlibrarians. Disponível em: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html. Acesso: 20 out. 2019.

GONZÁLEZ ALCAÍDE, Gregório *et al*. La colaborácion científica como objeto de estúdio. *In*: _____. **La colaborácion científica**: uma aproximación multidisciplinar. Valencia: Nau Libres, 2013. P.13-16.

HARDING, Rachel J. **About LabScribbles**. Blog LabScribbles. 2016. Disponível em: https://labscribbles.com/about/. Acesso em: 02 nov. 2019.

HARDING, Rachel J. Open notebook science can maximize impact for rare disease projects. **Plos**, n. 28, jan. 2019. Disponível em https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000120. Acesso em: 21 maio 2015.

HASTINGS, Rachel. Linked Data in libraries: status and future direction. **Infotoday.com**, Medford, MA, USA, v. 35, n. 9, nov. 2015. Disponível em:

http://www.infotoday.com/cilmag/nov15/Hastings--Linked-Data-in-Libraries.shtml. Acesso em: 16 jun. 2017.

HENNING, Patrícia Corrêa et al. Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Pesq. Bras. em Ci. da Inf. e Bib**., João Pessoa, v. 14, n. 3, p. 175-192, 2019.

HEY, Tony; HEY, Jessie. E-Science and its implications for the library community. **Library Hi Tech**, 01 October 2006, v. 24, n. 4, 2006, pp.515-528. DOI 10.1108/07378830610715383

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. **The fourth paradigma**: data intensive scientific discovery. Washington: Microsoft Research Redmond, 2009. 252 p.

HEY, Tony; TREFETHEN, Anne. e-Science and its implications. **Phil. Trans. R. Soc. Lond**. The Royal Society, 25 jun. 2003, n.361, p.1809-1825.
https://scienceblogs.com/pontiff/2008/06/26/pseudo-open-notebook-science . Acesso em: 12 set. 2019.

HEY, Tony; TREFETHEN, Anne. E. The UK e-Science core programme and the Grid. **Future Generation Computer Systems**, v. 18, n. 8, p. 1017-1031, Oct. 2002. Disponível em: https://doi.org/10.1016/S0167-739X(02)00082-1. Acesso em: 12 set. 2019.

HIGGINS, Sarah. The DCC curation lifecycle model. **International Journal of Digital Curation**, v. 3, n.1, dec. 2008. DOI: https://doi.org/10.2218/ijdc.v3i1.48

HODSON, Simon *et al*. **Turning FAIR into reality**. European Commission. Luxembourg: Publications Office of the European Union, 2018. Disponível em: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Acesso em: 05 jan.2020.

IBICT. INSTITUTO BRASILEIRO DE INFORMAÇÃO E TECNOLOGIA. **Manifesto de acesso aberto a dados de pesquisa brasileira para Ciência Cidadã**. 2016. Texto digital. Disponível em: http://www.ibict.br/Sala-de-Imprensa/noticias/2016/ibict-lanca-manifesto-de-acesso-aberto-a-dados-da-pesquisa-brasileira-para-ciencia-cidada/#_ftn1. Acesso em: 7 dez. 2017.

IFLA. INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. **Functional requirements for bibliographic records**: final report. München: K. G. Saur, 1998. Disponível em: http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf. Acesso em: 15 abr. 2017.

IFLA. INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. **Declaración de principios internacionales de catalogación** (PIC). IFLA, 2016.

ISOTANI, S.; BITTENCOURT, Ig I. **Dados abertos conectados**. São Paulo: Novatec, 2015. Disponível em: http://ceweb.br/livros/dados-abertos-conectados/. Acesso em: 10 nov. 2019.

JESSOP, David M; ADAMS, Sam E; MURRAY-RUST, Peter. Mining chemical information from open patentes. **Journal of Cheminformatics**, v.3, n. 40. 2011. Disponível em: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-40. Acesso em: 20 jun. 2020.

KELLOGG, Gregg; CHAMPIN, Pierre-Antoine; LONGLEY, Dave. **JSON-LD 1.1**: a JSON-based serialization for linked data. 2020. Disponível em: https://www.w3.org/TR/2020/CR-json-ld11-20200316/. Acesso em: 01 abr. 2020.

KITCHENHAM, Barbara; CHARTERS, Stuart. **Guidelines for performing systematic literature reviews in software engineering**. Technical Report EBSE 2007-001, Keele University; Durham University Joint Report, 2007. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4108896/mod_resource/content/2/slrPCS5012_highlighted.pdf. Acesso em: 10 maio 2020.

KNORR-CETINA, K. **Epistemic cultures**: how the sciences make knowledge. Harvard University Press, 1999.

KUHN, Thomas S. **A estrutura das revoluções científicas**. 5. ed. São Paulo: Ed. Perspectiva, 1998. 257 p.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. Ed. Brasília, DF: Brinquet de Lemos/Livros, 2004. 452 p.

LATOUR, Bruno; WOOLGAR, Steve. **A vida de laboratório**: a construção de fatos científicos. Rio de Janeiro: Relume-Dumará, 1997.

LAUFER, Carlos. **Guia de web semântica**. São Paulo: CeWeb.br, 2015. Disponível em: https://nic.br/media/docs/publicacoes/13/Guia_Web_Semantica.pdf. Acesso em: 07 mar. 2020.

LÓSCIO, Bernadette F.; BURLE, Caroline; CALEGARI, Newton. **Data on the Web best practices**. W3C, 2017. Texto digital. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 20 mar. 2020.

LÓSCIO, Bernadette; BURLE, Caroline; OLIVEIRA, Marcelo I.; CALEGARI, Newton. **Fundamentos para publicação de dados na Web**. São Paulo: Comitê Gestor da Internet no Brasil, 2018. 64 p. Disponível em: https://ceweb.br/media/docs/publicacoes/1/fundamentos-publicacao-dados-web.pdf. Acesso em: 24 dez. 2019.

MARTELETO, Regina Maria. Análise de redes sociais: aplicação nos estudos de transferência da informação. **Ciência da Informação**, Brasília, DF, v. 30,n.1, p.71-81, jan./abr. 2001. Disponível em: http://www.scielo.br/pdf/ci/v30n1/a09v30n1.pdf. Acesso em: 01 dez. 2019.

METTLER TOLEDO. **Reações de polimerização**: entendimento abrangente de cinética para desenvolver química de polímero sintético. 2020. Disponível em: https://www.mt.com/br/pt/home/applications/L1_AutoChem_Applications/L2_ReactionAnalysis/L2_Polymerization.html#publications. Acesso em: 29 mar. 2020.

MEDEIROS, Jackson da Silva; CAREGNATO, Sônia Elisa. Compartilhamento de dados e e-Science: explorando um novo conceito para a comunicação científica. **Liinc em Revista**, v.8, n.2, setembro, 2012, Rio de Janeiro, p. 311-322.

MEADOWS, Arthur Jack. **A comunicação científica**. Brasília, DF: Brinquet de Lemos/Livros, 1999. 268 p.

MENDES, Luis Augusto Lobão. Redes de colaboração: o poder da colaboração em massa. **Revista DOM**, Minas Gerais, p. 95-105, 2009.

MEY, Eliane Serrão Alves; SILVEIRA, Naira Christofoletti. **Catalogação no plural**. Brasília, DF: Briquet de Lemos, 2009.

MICROSOFT AZURE. **O que é middleware**. 2020. Disponivel em: https://azure.microsoft.com/pt-br/overview/what-is-middleware/. Acesso em: 13 jan. 2020.

MURRAY-RUST, Peter. **Jean-Claude Bradley**: hero of open notebook science; it must become the central way of doing science. Petermr's blog: a Scientist and the Web (Blog). [S.l.: s.n.].Publicado em 19 mai. 2014. Disponível em: https://blogs.ch.cam.ac.uk/pmr/2014/05/19/jean-claude-bradley-hero-of-open-notebook-science-it-must-become-the-central-way-of-doing-science/ . Acesso em: 02 nov. 2019.

NATIONAL REARCH COUNCIL. **A question of balance**: private rights and the public interest in scientific and technical databases. Washington: NRC, 1999. Disponível em: https://www.nap.edu/read/9692/chapter/1#ii. Acesso em: 02 jun. 2018.

NATIONAL SCIENCE BOARD (NSB). **Long-lived digital data collections**: enabling research and education in the 21st century. Alexandria, USA: National Science Foundation, 2005. Disponível em: https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf. Acesso em: 2 jul. 2018.

NATIONAL SCIENCE FOUNDATION (NSF). Where discoveries begin. **Dissemination abd sharing of research results**. Alexandria, 2015. Disponível em: http://www.nsf.gov/bfa/dias/policy/dmp.jsp. Acesso em: 2 jul. 2018.

OLIVEIRA, Adriana Carla Silva de. **Desvendando a autoralidade colaborativa na e-Science sob a ótica dos direitos de propriedade intelectual**. Orientador: Guilherme Ataíde Dias. 2016. 297 f. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, 2016.

OLIVEIRA, Adriana Carla Silva de; SILVA, Edilene Maria. Ciência aberta: dimensões para um novo fazer científico. **Inf. Inf**., Londrina, v.21, n.2, p. 5-39, maio/ago, 2016.

OLIVEIRA, Paulo Jorge; RODRIGUES, Fátima; HENRIQUES, Pedro Rangel. **Limpeza de dados**: uma visão geral. Disponível em: http://wiki.di.uminho.pt/twiki/pub/Research/Doutoramentos/SDDI2004/ArtigoOliveira.pdf. Acesso em: 19 jun. 2020.

OLIVER, Chris. **Introdução à RDA**: um guia básico. Brasília, DF: Briquet de Lemos, 2011.

ONSN. OPEN NOTEBOOK SCIENCE NETWORK. **Why should you keep na open notebook?** 2019?. Disponível em: http://onsnetwork.org/what-is-open-notebook-science/why-should-you-keep-an-open-notebook/ . Acesso em: 21 set. 2019.

OPEN DEFINITION. **Conformant licenses**. Projeto da Open Knowledge Foudation. 2020. Disponível em: http://opendefinition.org/licenses/. Acesso em: 23 fev. 2020.

OPEN DEFINITION. **The open definition**. [2004?]. Disponível em: http://opendefinition.org/. Acesso em: 23 fev. 2020.

OPEN GRAPH PROTOCOL. **The graph protocol**. Disponível em: https://ogp.me/. Acesso em: 02 mar. 2020.

OPEN KNOWLEDGE INTERNATIONAL. **The open data handbook**. 2004. Disponível em: http://opendatahandbook.org/guide/en/what-is-open-data/. Acesso em: 24 set. 2017.

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **Principles and Guidelines for Access to Research Data from Public Funding**. OECD, 2007. Disponível em: http://www.oecd.org/sti/sci-tech/38500813.pdf. Acesso em: 10 nov. 2017.

PADRON, Marcos Fragmomeni. **Uma proposta de modelo conceitual para representação da música popular brasileira**. Orientador: Fernando William Cruz. 2019. 221 f. Dissertação (Mestrado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília, 2019.

PATRÍCIO, H. S. A Europeana e a agregação de metadados na web: análise dos esquemas ESE/EDM e da aplicação de standards da web semântica a dados de bibliotecas. *In*: Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, 11., 2012, Lisboa. **Anais eletrônicos**... Lisboa, 2012.

PIRES, M. T. **Guia de dados abertos**. São Paulo: NIC.br, 2015. Disponível em: http://ceweb.br/guias/dados-abertos/. Acesso em: 08 ago. 2018.

POPAY, J.; ROGERS, A.; WILLIAMS, G. Rationale and standards for the systematic review of qualitative literature in health services Research. **Qualitative Health Research**, v. 8, n. 3, p. 341-351. DOI: 10.1177/104973239800800305

POWER, Lucy. **e-Research in the life sciences**: from invisible to virtual colleges. 2012. Disponível em: https://ora.ox.ac.uk/objects/uuid:de32d659-8908-4ebe-ab50-3ba6330f456a Acesso em: 28 abr. 2018.

RAUPP, Fabiano Maury; BEUREN, Ilse Maria. Metodologia da pesquisa aplicável às ciências sociais. *In*: BEUREN, I. M. (org.). **Como elaborar trabalhos monográficos em contabilidade**: teoria e prática. 3. ed. São Paulo: Atlas, 2006. p. 76-97.

RAUTENBERG, Sandro; SOUZA, Lucélia de; DALL'AGNOL, Josiane M. H.; MICHELON, Gisane A. **Guia prático para publicação de dados abertos na Web**. Curitiba: Appris, 2018. 280 p.

RIBES, David; LEE, Charlotte. P. Sociotechnical studies of cyberinfrastructure and e-research: current themes and future trajectories. **Computer Supported Cooperative Work**, v. 19, n.3-4, p. 231-244, 2010. Disponível em: https://link.springer.com/content/pdf/10.1007/s10606-010-9120-0.pdf. Acesso em: 15 set. 2019.

RILEY, J. **Understanding metadata**: what is metadata, and what is it for? Baltimore, MD: NISO Primer, 2017. Disponível em:http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf. Acesso em: 02 jun. 2018.

RIVA, P.; LE BOEUF, P.; ŽUMER, M. **IFLA Library Reference Model**: definition of a conceptual reference model to provide a framework for the analysis of non-administrative metadata relating to library resources. Netherlands: IFLA, 2017. 101 p.

ROCHA, Lucas de Lima; SALES, Luana Faria; SAYÃO, Luís Fernando. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso**, Salvador, v.11, n.3, p. 2-16, dez. 2017

RYAN, C. *et al*. Linked data authority records for Irish place names. **In J Digit Libr**, Berlin Heidelberg, v. 15, p. 73-85, abr. 2015.

SALES, Luana Farias. **Integração semântica de publicações científicas e dados de pesquisa**: proposta de modelo de publicação ampliada para a área de ciências nucleares. Orientadores: Rosali Fernandez de Souza e Luís Fernando Sayão. 2014. 264 f. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Instituto Brasileiro de Informação e Tecnologia. Universidade Federal do Rio de Janeiro, 2014.

SANT'ANA, Ricardo César Gonçalves. 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 eletrônicos**...Florianópolis, 2013. Disponível em: http://enancib.ibict.br/index.php/enancib/xivenancib/paper/view/4383. Acesso em: 30 jan. 2020.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Inf. Inf.**, 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. Acesso em: 30 jan. 2020.

SANTAREM SEGUNDO, José Eduardo. **Representação iterativa**: um modelo para repositórios digitais. Orientadora: Silvana Aparecida Borsetti Vidotti. 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

SANTAREM SEGUNDO, Jose Eduardo. Tim Berners-Lee e a ciência da informação: do hipertexto à Web Semântica. *In*: _____; SILVA, Márcia Regina da; MOSTAFA, Solange Puntel (Orgs.). **Os pensadores e a Ciência da Informação**. Rio de Janeiro: E-papers, 2012. p. 101-109.

SANTAREM SEGUNDO, Jose Eduardo. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16. 2015. **Anais eletrônicos**... [*S.l.: s.n.]*, 2015. Disponível em: http://www.ufpb.br/evento/index.php/enancib2015/enancib2015/paper/viewFile/3149/1193. Acesso em: 17 abr. 2016.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a recuperação de dados usando SPARQL. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO: além das nuvens, expandindo as fronteiras da Ciência da Informação, 15., 2014. Belo Horizonte. **Anais eletrônicos**... Belo Horizonte: UFMG/ ECI, 2014. p. 3863-3882. Disponível: http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/3191/2014_GT8-CO_09.pdf?sequence=1. Acesso em: 20 abr. 2018.

SANTAREM SEGUNDO, José Eduardo. Web semântica: fluxo para publicação de dados abertos e ligados. **Informação em Pauta**, Fortaleza, v. 3, número especial, p. 117-140, nov. 2018. DOI:

https://doi.org/10.32810/2525- 3468.ip.v3iEspecial.2018.39721.117-140. Acesso em: 01 jan. 2019.

SANTOS, Plácida Leopoldina Ventura Amorim da Costa; PEREIRA, Ana Maria. **Catalogação**: breve história e contemporaneidade. Niterói: Intertexto, 2014.

SANTOS, Plácida Leopoldina Ventura Amorim da Costa; CORRÊA, Rosa Maria Rodrigues. **Catalogação**: trajetória para um código internacional. Niterói: Intertexto, 2009.

SANTOS, P. L. V. A. da C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação**, v. 42, n. 2, p. 199-209, maio/ago. 2013. Disponível em: http://revista.ibict.br/ciinf/article/view/1382 . Acesso em: 12 dez. 2017.

SANTOS, Paulo Roberto; BORGES, Renata Silva; LORENÇO, Francisco dos Santos. Documentos de arquivo produzidos pela atividade científica: uma análise dos cadernos de laboratório do Instituto Oswaldo Cruz. Hist. Cienc. Saúde-Manguinhos, v.26, n.3, Rio de Janeiro. 2019.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**. Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015.

SCHAPIRA, Matthieu; HARDING, Ranchel J. Open laboratory notebooks: good for Science, good for society, good for scientists. **F1000Research Open for Science**, 2019. Disponível em: https://doi.org/10.12688/f1000research.17710.1. Acesso em: 21 set. 2019.

SCHIERMEIER, Quirin. 'You never said my peer review was confidential' – scientist challenges publisher. **Nature**, 24 jan. 2017. n.541, v.446. Disponível em: doi : 10.1038 / nature.2017.21342. Acesso em: 12 out. 2019.

SCHELL, Santiago. Tem Simple Rules for a computational biologist's laboratory notebook. **PloS Computational Biology**. São Francisco, v.11, n.9, 2015. Disponível em: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004385. Acesso em: 25 fev. 2017.

SEMELER, Alexandre Ridas. **Ciência da Informação em contextos de e-Science**: bibliotecários de dados em tempos de data Science. Orientador: Adilson Luiz Pinto. 2017. 164 f. Tese (Doutorado) - Programa de Pós-Graduação em Ciência da Informação do Centro de Ciências da Educação, Universidade Federal de Santa Catarina, 2017.

SERRA, Liliana Guisti. **A Web Semântica na gestão de livros digitais licenciados**: uma proposta de modelo. Orientador: José Eduardo Santarem Segundo. 2019. 155f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho, 2019.

SHAPIN, Steven; SCHAFFER, Simon. **Leviathan and the air-pump**: Hobbes, Boyle and the experimental life. Princeton University Press, 1985.

SILVA, Fabiano Couto Corrêa da. **Gestão de dados científicos**. Rio de Janeiro: Interciência, 2019. 128 p.

SILVA, Juliana Rocha de Faria. **Diretrizes para organização de informação musical brasileira**. Orientador: Fernando William Cruz. 2017. 287 f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília, 2017.

SILVA, Luciana Candida da; SANTAREM SEGUNDO, José Eduardo; SILVA, Marcel Ferrante. Princípios FAIR e melhores práticas do Linked Data na publicação de dados de pesquisa. **Informação&Tecnologia** (ITEC), Marília/João Pessoa, v.5, n.2, p.81-103, jul./dez. 2018. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/44812/27746. Acesso em: 07 mar. 2020.

SILVA, Luciana Candida da; SANTAREM SEGUNDO, José Eduardo; ZAFALON, Zaira Regina; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. O código RDA e a iniciativa BIBFRAME: tendências da representação da informação no domínio bibliográfico. **Em Questão**, Porto Alegre, v. 23, n. 3, p. 130-156, set./dez. 2017. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/69549/41062. Acesso em: 20 jan. 2020.

SIMIONATO, A. C. Mapeamento dos metadados para dados científicos. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18. 2017. Marília, SP: UNESP, 2017. Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/563/874>. Acesso em: 6 jun. 2018.

SONNENWALD, Diane H. Scientific collaboration. *In*: CRONIN, B. **Annual Review of Information Science and Technology**. Medford: Information Today, 2007. v.41, cap.14, p. 643-681.

STRASSER, Carly *et al*. **DataONE**: primer on data management: what you Always wanted to know. 2012. Disponível em: https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf. Acesso em: 21 dez. 2019.

TAYLOR, Arlene G. **The organization of the information**. Westport: Libraries Unlimited, 2003. 280 p.

TEIXEIRA, Marcelo Votto. **Curso RDA**. ClassCursos. 2019.

TENOPIR, C. *et al*. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. **PLoS One**, v. 10, n.8, 2015. Disponível em: DOI 10.1371/journal.pone.0134826. Acesso em: 02 set. 2019.

TENOPIR, C. *et al*. Data sharing by scientists: practices and perceptions. **Plos One**, v.6, n.6, 2011. Estados Unidos. Disponível em: https://doi.org/10.1371/journal.pone.0021101. Acesso em: 21 set. 2019.

THE WORLD BANK. **Open government data toolkit**. 2013. Disponível em: <http://opendatatoolkit.worldbank.org/en/index.html>.Acesso em: 15 jan. 2020.

TILLETT, Bárbara. **FRBR and RDA**: resource description and access. *In*: TAYLOR, A. G.

Understanding FRBR: what it is and how it will affect our retrieval tools. Westport, Ct: Greenwood Publishing Group, 2007, p. 87-95.

TODD, Matthew. **Open source drug discovery for malaria**. The Synaptic Leap open source biomedical research (Blog). Publicado em 25 jul. 2011. Disponível em: http://www.thesynapticleap.org/node/343. Acesso em: 12 set. 2019.

TOMAÉL, Maria Inês. Redes de conhecimento. **DataGramaZero**. Revista de Ciência da Informação, Rio de Janeiro, v.9, n.2, p.1-13, abr. 2008.

TORRES-SALINAS, D.; ROBINSON-G\ARCÍA, N.; CABEZAS-CLAVIJO, A. Compartir los datos de investigación em ciência: introducción al data sharing. **El professional de la información**, v. 21, n.2, p. 173-184, mar./abr. 2012.

UNESP. Programa de Pós-Graduação em Ciência da Informação. **Linhas de Pesquisa**. 2019. Disponível em: https://www.marilia.unesp.br/#!/pos-graduacao/mestrado-e-doutorado/ciencia-da-informacao/linhas-de-pesquisa/. Acesso em: 04 abr. 2019.

VAZ, Glauber José. **E-Science na Emprapa**. Campinas, SP: Embrapa Informática Agropecuária, 2011. 58 p. Série Documentos, n. 117.

VERHAAR, Peter. **Report on object models and functionalities**. DRIVER II, 2008.

W3C. World Wide Web Consortium. **JSON-LD 1.1**: uma serialização baseada em JSON para dados vinculados. Recomendação para candidatos do W3C 16 de março de 2020. Disponível em: https://www.w3.org/TR/json-ld/. Acesso: 24 mar. 2020.

W3C. World Wide Web Consortium. **XML technology**. 2015. Disponível em: https://www.w3.org/standards/xml/. Acesso em: 27 mar. 2020.

W3C. World Wide Web Consortium. **RDF - Resource Description Framework**. 2014. Disponível em: https://www.w3.org/RDF/. Acesso em: 27 mar. 2020.

W3C. World Wide Web Consortium. **SPARQL Query Language for RDF**. 2008. Disponível em: https://www.w3.org/TR/rdf-sparql-query/. Acesso em: 20 out. 2020.

WELSH, Anne; BATLEY, Sue. **Practical cataloguing**: AACR2, RDA and MARC21. [S.l.]: ALA Neal-Schuman, 2012. cap. 1, p. 1-6.

WIGGINS, Andrea *et al*. **Data management guide for public participation in scientific research**. DataONE Public Participation in Scientific Research Working Group. Fev. 2013. Disponível em: https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf . Acesso em: 12 jan. 2020.

WILKINSON, M. *et al*. The FAIR guiding principles for scientific data management and stewardship. **Sci Data**, n. 3, 2016. DOI: 10.1038/sdata.2016.18

WILLIGHAGEN, Egon; BRANDLE, Martin. Resource description framework technologies in chemistry. **Journal of Cheminformatics** v.5, n.15. 2011. Diponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118380/. Acesso em: 15 jun. 2020.

YANG, Xiaoyu; WANG, Lizhe; LASZEWSKI, Gregor Von. Recent Research Advances in e-Science. **Cluster Comput**, n.12, set. 2009, pp. 353–356. DOI 10.1007/s10586-009-0104-0.

ZHU, Yimei; PROCTER, Rob . Use of blogs, Twitter and Facebook by UK PhD students for scholarly communication. **Observatorio (OBS*) Journal**, v.9, n.2, 2015.