UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

MARCELLO MUNDIM RODRIGUES

REPOSITÓRIOS DE DADOS CIENTÍFICOS NA AMÉRICA DO SUL: uma análise da conformidade com os Princípios FAIR

BELO HORIZONTE/MG
2020

MARCELLO MUNDIM RODRIGUES

REPOSITÓRIOS DE DADOS CIENTÍFICOS NA AMÉRICA DO SUL: uma análise da conformidade com os Princípios FAIR

Dissertação apresentada como requisito à defesa e conclusão de Mestrado em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação, Universidade Federal de Minas Gerais (ECI-UFMG).
Linha de pesquisa: Gestão e Tecnologia em Informação e Comunicação (GETIC).
Orientadora: Cíntia de Azevedo Lourenço.
Coorientador: Guilherme Ataíde Dias.

BELO HORIZONTE/MG

2020

Dados Internacionais de Catalogação na Publicação (CIP)

R696r
2020

Rodrigues, Marcello Mundim, 1985
Repositórios de dados científicos na América do Sul: uma análise da conformidade com os Princípios FAIR. / Marcello Mundim Rodrigues. - Belo Horizonte, 2020.
110 f. : il.

Orientadora: Cíntia de Azevedo Lourenço.
Coorientador: Guilherme Ataíde Dias.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Programa de Pós-graduação em Gestão e Organização do Conhecimento.
Inclui bibliografia.

1. Repositórios digitais. 2. Dados científicos – Gestão e curadoria. 3. Princípios FAIR. 4. América do Sul – Repositórios de dados. I. Título. II. Lourenço, Cíntia de Azevedo. III. Dias, Guilherme Ataíde. IV. Universidade Federal de Minas Gerais. Programa de Pós-graduação em Gestão e Organização do Conhecimento.

CDU: 659.2

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

FOLHA DE APROVAÇÃO

REPOSITÓRIOS DE DADOS CIENTÍFICOS NA AMÉRICA DO SUL: uma análise da conformidade com os Princípios FAIR

MARCELLO MUNDIM RODRIGUES

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia.

Aprovada em 28 de agosto de 2020, pela banca constituída pelos membros:

Prof(a). Cíntia de Azevedo Lourenço (Orientadora)
ECI/UFMG [por videoconferência]

Prof(a). Guilherme Ataíde Dias (Coorientador)
UFPB [por videoconferência]

Prof(a). Fabiano Couto Corrêa da Silva
UFRGS [por videoconferência]

Prof(a). Carlos Henrique Marcondes de Almeida
UFF [por videoconferência]

Belo Horizonte, 28 de agosto de 2020.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

ATA DA DEFESA DA DISSERTAÇÃO DO ALUNO **MARCELLO MUNDIM RODRIGUES**

Realizou-se, no dia 28 de agosto de 2020, às 14:00 horas, Videoconferência, da Universidade Federal de Minas Gerais, a defesa de dissertação, intitulada *REPOSITÓRIOS DE DADOS CIENTÍFICOS NA AMÉRICA DO SUL: uma análise da conformidade com os Princípios FAIR*, apresentada por MARCELLO MUNDIM RODRIGUES, por videoconferência, número de registro 2018666902, graduado no curso de BIBLIOTECONOMIA/DIURNO, como requisito parcial para a obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Cíntia de Azevedo Lourenço - ECI/UFMG [por videoconferência] (Orientadora), Prof(a). Guilherme Ataíde Dias - UFPB [por videoconferência] (Coorientador), Prof(a). Fabiano Couto Corrêa da Silva - UFRGS [por videoconferência], Prof(a). Carlos Henrique Marcondes de Almeida - UFF [por videoconferência].

A Comissão considerou a dissertação:

(X) Aprovada

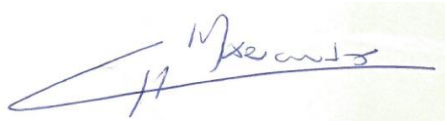
( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 28 de agosto de 2020.

Prof(a). Cíntia de Azevedo Lourenço

Prof(a). Guilherme Ataíde Dias

Prof(a). Fabiano Couto Corrêa da Silva



Prof(a). Carlos Henrique Marcondes de Almeida

À pessoa que me ensinou a ter fé na vida, a perseverar, a acreditar em dias melhores.

Àquela que foi um exemplo de humildade e alegria. Minha segunda mãe, vó Julia Mundim.

AGRADECIMENTOS

Aparentemente, esse parece ser o fim de uma jornada, um vislumbre da minha Torre. Ávido pelo fim, deixei de olhar com carinho para a estrada. Por vezes pensei em mudar o caminho, mas havia força, e por ela fui impelido. A ela dou nomes, pois mortal que sou, por meio dessa posso imortalizá-los.

Agradeço à pessoa mais importante da vida, pois é a origem de tudo, minha mãe, Eliane Maria Mundim Neto. Você merece todo crédito. Agradeço a meu irmão, Marcus Vinícius Mundim Rodrigues, pela parceria e prova de resiliência.

Agradeço a meus avós, Antônio Neto Junior e Julia Mundim Neto, por tudo e mais, pois foram prova de amor incondicional. Eternamente em nossos corações.

Agradeço a minha madrinha, Janine de Fátima Mundim Neto, por todo apoio, incentivo, carinho e cuidado. Pelos livros, pelas conversas, por acreditar em mim.

Agradeço a minhas tias-avós de Paracatu, Dió, Guili e Loló, pois sem elas a vida não teria graça.

Agradeço a meus orientadores, professores Cíntia de Azevedo Lourenço e Guilherme Ataíde Dias, por todo ensinamento e paciência.

Agradeço a meus tios e tias, Eneida do Carmo Mundim Neto, Eric Alberto Lima de Oliveira, Hélder Magela Mundim Neto e Paula Cristina Barboza de Matos Mundim, Jane Magnólia Mundim Neto, Janete Aparecida Mundim, Wander Mundim Neto e Clean Santos de Oliveira Mundim, pelo carinho, cuidado e bom exemplo.

Agradeço a meus primos e amigos, especialmente, Hélder Filho Mundim e Príscia Laíse de Matos Mundim, Vinícius Mateus Mundim, Felipe Mundim Nunes, Guilherme André Santana e Leonardo André Mendes, Marcos Duarte Guimarães Filho, Pedro Pimentel dos Santos, Wagner Mesquita e Ana Luiza Maggi Marcatto, pois vocês são únicos e fizeram toda a diferença.

É sabido que um homem sem ninguém é um homem sem história e, portanto, considero a minha rica e cheia de personagens apaixonantes.

Obrigado por tudo!

III

Se ao seu conselho eu devesse me desviar
Para aquele curso sinistro que, é sabido,
Esconde a Torre Negra? Porém eu, de boa-fé imbuído,
Tomei o indicado caminho, sem orgulho demonstrar
Nem esperança rediviva ao ver o fim se aproximar,
Mas sim gratidão pela ideia de algum fim existir.

XXXIV

Ali estavam eles, pelos lados dos montes, unidos
Para assistir meu fim. Eu, uma moldura animada
Para mais um quadro! Numa súbita labareda
Eu os vi e reconheci a todos. E, destemido,
Deixei meus lábios formarem um bramido:
“Childe Roland à Torre Negra chegou”.

(ROBERT BROWNING, 1855)

Nós estamos normalmente mais assustados
do que machucados; e sofremos mais na
imaginação do que na realidade.

(SÊNECA)

RESUMO

O crescimento do processo de digitalização no mundo tem apresentado desafios à prática bibliotecária, uma vez que tradicionalmente seu núcleo de atuação está na organização de documentos bibliográficos em formato físico. A intenção de pesquisa teve como fim estudar o fenômeno dos dados gerados por meio do processo científico e o desenvolvimento de serviços que enfrentam os crescentes desafios de sua gestão e curadoria, o que envolve volumes de recursos digitais em constante expansão. O problema de pesquisa se encontra nos ambientes e nas práticas responsáveis pela organização desses ativos digitais resultantes da investigação científica contemporânea. Foram objetos de estudo dessa investigação: os dados, enquanto fenômeno; os conjuntos de dados, enquanto unidades informacionais; os Princípios FAIR, enquanto diretrizes à gestão e curadoria de dados científicos; e os repositórios digitais institucionais de dados científicos, enquanto ambientes virtuais de organização informacional. As questões de pesquisa respondidas foram: Qual a natureza dos conjuntos de dados científicos arquivados em repositórios digitais institucionais oriundos do continente sul-americano? Qual a conformidade desses repositórios com os Princípios FAIR? O objetivo da pesquisa foi investigar conjuntos de dados científicos e respectivos repositórios digitais institucionais sul-americanos à luz dos Princípios FAIR. Lidar com o fenômeno dos dados é assistir a comunidade acadêmica em seus diversos campos, assim como a sociedade civil em suas dificuldades tecnológicas e digitais, disponibilizando meios de acesso mais ágeis e assertivos num ambiente digital cada vez mais nebuloso e caótico. A investigação consistiu em uma pesquisa aplicada, do ponto de vista de sua natureza; de método qualitativo, pelos meios de sua abordagem ao problema; exploratória e analítica, do ponto de vista dos objetivos; bibliográfica e documental, a partir dos procedimentos técnicos. Fez-se levantamento dos repositórios de dados científicos na base RE3DATA. A coleta dos dados contou com levantamento bibliográfico em bases de dados científicas, e com levantamento de dados referentes aos depósitos de conjuntos de dados científicos em repositórios vinculados a instituições sul-americanas. Utilizou-se da análise de conteúdo à concepção dos resultados de pesquisa. Os repositórios encontrados e aptos à investigação foram o do PPBio (Brasil); o da UFPR (Brasil); o do CIAT (Colômbia); o da PUC (Peru); o do MEC (Peru); e o da USIL (Peru). Os achados indicam que os programas (*software*) responsáveis pelos repositórios investigados que servem à gestão e curadoria de dados científicos são o *Morpho* (*DataONE*), o *DSpace*, e o *Dataverse*. Em relação aos conjuntos de dados científicos investigados, notou-se que sua natureza se concentra em dados textuais e numéricos, salvos em arquivos de texto e em tabelas, respectivamente. Os repositórios em maior conformidade com os Princípios FAIR foram aqueles estabelecidos mediante o uso do *Dataverse*. Concluiu-se que profissionais da informação devem buscar sua capacitação em dados, a começar pelo planejamento de projetos e políticas institucionais dirigidas à implementação de repositórios de dados científicos, passando pelo entendimento das divergentes necessidades entre comunidades, pelo conhecimento técnico computacional exigido a tais práticas, e idealmente, pela busca da padronização e manutenção desses serviços.

Palavras-chave: Repositórios digitais. Dados científicos – Gestão e curadoria. Princípios FAIR. América do Sul – Repositórios de dados.

ABSTRACT

The growth of the digitization process in the world has presented challenges to the librarian practice, once traditionally its acting core is in the organization of bibliographic documents. The research had as an end studying the data phenomenon generated by the scientific process and the development of services that face the rising challenges of data management and curation, which involves volumes of digital resources in constant expansion. The research problem is on the environments and practices responsible for digital asset organization resulting from the contemporary scientific investigation. The study objects of such inquiry were: the data, as a phenomenon; the datasets, as informational unities; the FAIR Principles, as guidelines to the scientific data management and curation; and the institutional digital repositories of scientific data, as virtual environments of informational organization. The research questions answered were: What is the nature of the scientific datasets archived in the south american institutional digital repositories? What is the compliance of such repositories with the FAIR Principles? The research objective was to investigate scientific datasets and respective south american institutional digital repositories in light of the FAIR Principles. Dealing with the data phenomenon is assisting the academic community in its diverse fields, as well as the civil society in its digital and technological difficulties, providing more agile and assertive means of access in an increasingly nebulous and chaotic digital environment. The investigation consisted of an applied research, from the point of view of its nature; of a qualitative method research, by means of its approach to the problem; of an exploratory and analytical research, from the point of view of the objectives; of a bibliographic and documentary research, based on technical procedures. The scientific data repositories were surveyed in the RE3DATA database. The data collection was made by surveying articles in scientific databases, and by surveying data related to scientific datasets deposited in repositories linked to south american institutions. Content analysis was used to obtain the research results. The repositories found and suitable for investigation were those created and managed by the following institutions: PPBio (Brazil); UFPR (Brazil); CIAT (Colombia); PUC (Peru); MEC (Peru); and USIL (Peru). The findings indicate that the software behind the investigated repositories which are fit to management and curation of scientific data are Morpho (DataONE), DSpace, and Dataverse. Regarding the investigated scientific datasets, it was noted that their nature is clustered in textual and numerical data, saved in text files and tables, respectively. The repositories in greater compliance with the FAIR Principles were established by the use of Dataverse. It was concluded that information professionals should seek their training in data, starting with the planning of projects and institutional policies aimed at the implementation of scientific data repositories, including the understanding of the divergent needs among communities, the technical computational knowledge required for such practices, and ideally, the search for standardization and maintenance of these services.

Keywords: Digital repositories. Scientific data – Management and curation. FAIR Principles. South America – Data repositories.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Dentro do processo evolutivo da ciência, observaram-se períodos na história humana que se destacaram pela maneira que a prática científica foi conduzida. Num primeiro momento, foram feitos experimentos e observações sobre o comportamento natural das coisas do mundo físico e passíveis de análise, o que se denominou ciência empírica. O empirismo trabalha variáveis distintas, em ambientes controlados ou não, e que busca validar ou refutar correlações, como causa-efeito.

Como consequência da experiência empírica, surgiu então um paradigma apoiado na observação e experimentação com a pretensão de testar hipóteses. Assim, hipóteses, teorias e leis foram fruto desse processo científico que se manteve e perdurou por séculos, até sofrer modificações a partir de meados do século XX.

Desde então, a humanidade usufrui da computação para gerar simulações e análises de dados em seus experimentos, observações e testes de hipóteses, contando inicialmente com o uso de computadores robustos e com baixas capacidades de armazenamento e processamento.

Com o passar das décadas, essas máquinas tiveram sua capacidade de processamento melhorada de forma a assistir mais efetivamente os pesquisadores em seu fazer diário, melhorando o tempo gasto e a qualidade da análise dos dados coletados.

Esse contexto histórico, conhecido pelo surgimento e evolução dos computadores e seus serviços, destaca nomes como Alan Turing, famoso pai da computação; Tim Berners-Lee, responsável pela *World Wide Web*; Bill Gates, fundador da empresa americana *Microsoft*; entre outros.

É a partir de meados da década de 1940, com o cenário do fim da segunda guerra mundial e começo da guerra fria, que o mundo observou o início de uma corrida armamentista e tecnológica, além de disputas de influências e de territórios entre os Estados Unidos e a antiga União Soviética, o que impactou na velocidade do desenvolvimento computacional e industrial.

Surgem também outros avanços em Ciência e Tecnologia (C&T) no período pós-segunda guerra, tais como pesquisas no uso de energia nuclear, a terceira revolução industrial, as Tecnologias de Informação e Comunicação (TICs), entre elas a *Web* e a *Internet*, ou seja, serviços estratégicos de inteligência numa disputa de poder entre potências.

A computação nesse cenário tem grande influência na aceleração do desenvolvimento da C&T em nível global, uma vez que se torna o pilar dos produtos e serviços da segunda fase do século XX. Por consequência e com a mesma velocidade, a informação

e a comunicação se tornam menos analógicas, adentrando, ao final de um milênio, na era digital.

Com o crescimento significativo do volume de dados ao longo dos anos de democratização tecnológica mundial, fase conhecida como globalização, aumenta a busca pelo desenvolvimento de poder de processamento e de armazenamento das máquinas frente à expansão do fenômeno dos dados e da informação. O termo *Big Data* surge nesse contexto, o qual tem influência de outro termo cunhado por Weinberg em 1961 que definia o crescimento da pesquisa científica no mundo, o *Big Science*[1].

Novos desafios às técnicas de tratamento dos dados e da informação emergem por conta do aumento exponencial do volume informacional. Essa se torna uma oportunidade para a Biblioteconomia e a Ciência da Informação buscarem abordar questões impostas pela digitalização e aumento do acesso à informação.

As bibliotecas virtuais e digitais surgem como respostas às necessidades de acesso remoto a serviços e produtos de bibliotecas tradicionais, sendo consequência da premissa bibliotecária de acesso universal ao conhecimento humano.

Em um primeiro momento, discutiu-se o papel da biblioteca tradicional em um mundo exclusivamente digital. Dessa forma, a instituição “biblioteca” foi colocada em xeque por muitos teóricos. Todavia, o que se percebe nos dias atuais são ambientes onde bibliotecas tradicionais, virtuais e digitais coexistem em uníssono, tendo foco na finalidade de sua existência, que consiste em assistir o usuário-cliente a satisfazer sua necessidade informacional.

Portanto, é nesse contexto em que o atual trabalho se encaixa, a partir do ponto em que reflete o lugar das bibliotecas em um universo em expansão, mais especificamente dentro do propósito acadêmico, em que se observa o desenrolar de uma fase que busca tratar tecnicamente o material científico produzido em laboratórios de pesquisa de programas de pós-graduação de universidades e institutos, assim como de instituições e agências de pesquisa científica, públicas e privadas.

Assim, bibliotecas universitárias e/ou especializadas que desejam salvaguardar, organizar e dar acesso à produção científica de sua instituição por meio digital devem desenvolver ambientes virtuais conhecidos como repositórios digitais institucionais. Esses repositórios podem ser classificados em documentais, bibliográficos, de dados, ou híbridos. Servindo ao propósito dessa investigação, classificam-se os dados em pessoais, institucionais, governamentais, e científicos.

[1] Grandes investimentos e volumes de dados gerados na pesquisa científica.

1.1 Problematização

O crescimento dos processos de digitalização e digitização no mundo tem apresentado desafios à prática bibliotecária, uma vez que tradicionalmente seu núcleo de atuação está na organização de documentos bibliográficos em formato físico. Com o aumento da atividade humana em ambientes digitais, cresce o número de documentos que são digitalizados (cópias) e que nascem nesse formato (nato digitais), que por sua vez necessitam de adequada gestão documental na busca por sua preservação, organização e recuperação. Dessa maneira, surge a preservação digital como campo de atuação e pesquisa científica na Biblioteconomia e Ciência da Informação.

Fundamentada na atividade bibliotecária, a preservação digital busca salvaguardar, organizar, disponibilizar, compartilhar, disseminar e permutar ativos digitais que compõem e dão sustento ao conhecimento humano.

Vários são os recursos atualmente disponíveis que procuram reduzir os problemas de organização e acesso a documentos e informações em ambientes digitais, como os esquemas ou padrões de metadados, os quais possibilitam a universalização (padronização) da descrição do conteúdo de objetos informacionais tendo como fim a recuperação da informação, os relacionamentos e a importação de registros (interoperabilidade) de um sistema a outro, pois padrões universais de linguagem e codificação podem ser reconhecidos por sistemas inteligentes. Esses recursos servem aos propósitos de acervos digitais em ascensão como os repositórios.

A intenção de pesquisa teve como fim estudar o fenômeno dos dados gerados por meio do processo científico e o desenvolvimento de serviços que enfrentam os crescentes desafios de sua gestão e curadoria, o que envolve volumes de recursos digitais em constante expansão. O problema de pesquisa se encontra nos ambientes e nas práticas responsáveis pela organização desses ativos digitais resultantes da investigação científica contemporânea.

Ao catalogar livros, um bibliotecário deve ter conhecimento de técnicas e códigos para que uma determinada obra e respectivo item[2] sejam apropriadamente incorporados ao acervo de sua coleção via *sotfware* de gestão de acervos, assim como também deve ter noção mínima dos assuntos abordados em seu texto para então indexá-los no sistema, almejando sua recuperação no futuro.

No contexto dos dados, armazenar, descrever, organizar, tratar e preservar informação pode ser bastante complexo devido a inúmeros fatores que cercam o desenvolvimento de uma pesquisa científica e seus resultados. Costumeiramente, os dados

[2] “Obra” se refere ao conhecimento contido em um livro, portanto, de natureza abstrata. “Item” se refere ao objeto físico impresso (unidade) que possui o conteúdo de uma obra, assim, de natureza concreta.

não são publicados, mas numa situação em que seja proposto fazê-lo, seria necessário obter e publicar metadados quanto ao: contexto da pesquisa e dos dados; processo de coleta e análise dos dados (descrição dos processos, técnicas e/ou *software* para replicação dos resultados); financiamento (que envolve direitos autorais e licenças para acesso e reúso); plano de gestão e ciclo de vida dos dados; entre outros.

Dessa forma, busca-se obter conhecimento da realidade acerca da disposição dos dados em ambientes virtuais, assim como da qualidade da descrição dos dados nos metadados utilizados por meio da gestão técnica desses repositórios.

Além disso, busca-se conhecer a natureza do conteúdo dos dados arquivados em repositórios digitais, e consequentemente, os formatos e extensões de arquivos que compõem os conjuntos de dados gerados ao longo de um processo científico. Em recente publicação, os pesquisadores Sales e Sayão (2019) apresentam uma taxonomia (Figura 1) que se refere à natureza dos dados científicos.

Figura 1 – Taxonomia quanto à natureza dos dados.

Fonte: SALES; SAYÃO, 2019, p. 41.

Para que o relato da pesquisa fique minimamente padronizado, decidiu-se por utilizar um termo dentre alguns que se referem a um dos objetos estudados. A literatura estrangeira comumente se direciona aos dados produzidos em meio científico como *scientific research data* ou apenas *research data.* Na brasileira, os termos mais utilizados por pesquisadores estão entre “dados de pesquisa” e “dados científicos”. Outros termos como “dados acadêmicos” ou “dados de pesquisa científica” também aparecem, porém em menor proporção. Há controvérsias quanto ao termo “pesquisa”, que pode envolver outras para além

dos laboratórios científicos, como a pesquisa de satisfação ou senso, ambas se tratando de levantamentos de dados (*surveys*). À vista disso, para que não haja tal confusão e de forma a simplificar a comunicação, adotar-se-á o termo “dados científicos” no texto, assim como na tradução de outros termos referentes a ele, salvo títulos de repositórios, citações diretas e suas respectivas referências.

Por dados científicos, entende-se que são as menores unidades informacionais (devido à natureza granular) coletadas, preservadas em documentos de forma analógica ou digital, estruturadas e analisadas por métodos científicos, que têm seu fim na produção e manutenção do conhecimento. Esses dados brutos possuem complexidade tal que são, muitas das vezes, apenas compreendidos por pesquisadores, máquinas e programas que os coletaram, geraram, simularam ou processaram, o que acaba se tornando um problema aos profissionais atuantes no campo da preservação digital, geralmente fora do ambiente e do domínio da pesquisa.

Na longa cauda da ciência, a diversidade da natureza dos dados coletados e agrupados é refletida em um número plural de arquivos salvos em diferentes formatos e extensões, gerando assim um ou mais conjuntos de dados científicos. Esses conjuntos podem ser definidos então como uma compilação de arquivos digitais que são gerados no processo científico-investigativo, e que por sua vez possuem dados de natureza e conteúdo heterogêneo. Ou seja, a natureza dos dados está diretamente relacionada aos formatos e extensões dos arquivos digitais que os agrupam.

> Quanto à natureza dos dados – retrata a grande diversidade e heterogeneidade de tipos de dados que podem ser originados no ambiente de pesquisa em termos de formatos, mídias, suportes, expressões, arcabouço tecnológico, etc. (SALES; SAYÃO, 2019, p. 43-44).

O meio para que os dados científicos encontrem seu fim está na comunicação científica (a ponta do *iceberg*), que reporta, na maioria das vezes, resultados condensados por meio da publicação de artigos científicos em periódicos especializados, da apresentação de trabalhos em congressos e respectiva publicação em anais, entre outros. “[...] A publicação dos resultados da sua pesquisa, e a literatura publicada é apenas a ponta do *iceberg* de dados. [...] Por *iceberg* de dados quero dizer que há muitos dados que são coletados, mas não tratados ou publicados de forma sistemática”. (HEY; TANSLEY; TOLLE, 2009, p. xvii, tradução nossa). Ressalta-se a relevância da neutralidade do pesquisador ao analisar seus dados para que não haja mal uso, excessos, ou manipulação de resultados.

Durante todo o processo de investigação científica, pesquisadores se utilizam de diversos meios para documentar seus achados (o uso de cadernos de pesquisa para registro de anotações em meio analógico ou eletrônico, entre outros) e arquivar seus documentos contendo dados coletados (computadores, celulares, hds, nuvem, etc.). “No século XX, os

dados nos quais teorias científicas eram baseadas ficavam geralmente ocultos em *notebooks* científicos individuais ou, por alguns aspectos da 'grande ciência', armazenados em mídia magnética que eventualmente se torna ilegível". (HEY; TANSLEY; TOLLE, 2009, p. xi, tradução nossa). Logo, esses dados por vezes se perdem em meio à má gestão e organização ou ao fim de um ciclo investigativo.

> [...] os cadernos [...] muitas vezes ficam limitados às paredes dos laboratórios, acentuando uma cultura de segredo que, cada vez mais, precisa ser discutida e superada. Os dados que sustentam uma pesquisa [...] geralmente ficam adormecidos, armazenados em computadores ou mídias pessoais [...]. [...] cadernos de laboratório [...] contistuem a espinha dorsal da guarda de registros, gestão de dados, análises iniciais e interpretação de resultados em pesquisas (ROCHA; SALES; SAYÃO, 2017, p. 3).

Uma reflexão sobre o tratamento e organização das informações desses dados consiste na realidade de que esses geralmente se constituem em conjuntos de objetos informacionais, que integrados podem ser processados como um fundo arquivístico. Caso esse tratamento seja individualizado, como na organização da informação de dados bibliográficos, o relacionamento entre os objetos informacionais de uma determinada pesquisa pode se perder, prejudicando a integridade das informações contidas nesses conjuntos de dados gerados em determinada investigação científica.

Com base nessa reflexão e no reconhecimento da complexidade desses conjuntos de dados, enfrentam-se os desafios técnicos impostos a seu arquivamento, descrição, organização, preservação, recuperação e reúso durante o desenvolvimento dessa pesquisa.

Portanto, são objetos de estudo dessa investigação: os dados, enquanto fenômeno; os conjuntos de dados, enquanto unidades informacionais; os Princípios FAIR, enquanto diretrizes à gestão e curadoria de dados científicos; e os repositórios digitais institucionais de dados científicos, enquanto ambientes virtuais de organização informacional.

Sendo assim, no contexto da pesquisa científica universal, bem como dos repositórios digitais que se propõem a arquivar, descrever, organizar, preservar, recuperar e dar acesso a conjuntos de dados científicos, as perguntas que se pretende ao final responder são:

Qual a natureza dos conjuntos de dados científicos arquivados em repositórios digitais institucionais oriundos do continente sul-americano?

Qual a conformidade desses repositórios com os Princípios FAIR?

1.2 Objetivos

1.2.1 Objetivo geral

Investigar conjuntos de dados científicos e respectivos repositórios digitais institucionais sul-americanos à luz dos Princípios FAIR.

1.2.2 Objetivos específicos

A) Identificar os repositórios de dados científicos criados e geridos por instituições de ensino superior e/ou agências de pesquisa e fomento sul-americanas;
B) Identificar e descrever os programas (*software*) responsáveis pelos repositórios identificados;
C) Identificar e descrever os formatos e extensões dos arquivos que compõem os conjuntos de dados científicos depositados nesses repositórios;
D) Verificar e comparar a conformidade desses repositórios de dados científicos com os Princípios FAIR.

1.3 Justificativa

Sabe-se que a Ciência da Informação (CI) “é, por natureza, interdisciplinar, embora suas relações com outras disciplinas estejam mudando” (SARACEVIC, 1996, p. 42), e que tem como proposta mor a organização do conhecimento humano. Logo, lidar com o fenômeno dos dados sob essa ótica é assistir a comunidade acadêmica em seus diversos campos, assim como a sociedade civil em suas dificuldades tecnológicas e digitais, disponibilizando meios de acesso ágeis e assertivos num ambiente digital cada vez mais nebuloso e caótico.

> A interdisciplinaridade foi introduzida na CI pela própria variedade da formação de todas as pessoas que se ocuparam com problemas descritos. Entre os pioneiros havia engenheiros, bibliotecários, químicos, linguistas, filósofos, psicólogos, matemáticos, cientistas da computação, homens de negócios e outros vindos de diferentes profissões ou ciências. [...] essa multiplicidade foi responsável pela introdução e permanência do objetivo interdisciplinar na CI (SARACEVIC, 1996, p. 48).

O volume de dados e informações a que se tem acesso atualmente se tornou imensurável, fato que traz consigo o desafio da escolha entre o que pode ser útil e o que se torna dispensável. É papel da CI buscar soluções práticas para esses casos, ou seja, facilitar e orientar o acesso, dar precisão e autonomia aos usuários e agentes informacionais na era digital e dos dados.

Um dos deveres do profissional da informação nesse crescente contexto é pôr em prática os princípios do acesso aberto (como prega a Biblioteconomia clássica em relação à democratização do conhecimento), que busca expandir o uso dos recursos científicos de forma a agilizar e ampliar o processo de construção do saber.

Esse movimento de abertura também equilibra a desigualdade elitista científica e dá retorno dos investimentos de ordem pública à sociedade, uma vez que se passa a exigir de pesquisadores acesso amplo a resultados e dados de pesquisas científicas financiadas pelo Estado.

Para que se possa tornar tais ideais realidade, é que se justifica o investimento de recursos em ambientes e normas que se destinam a desenvolver melhores práticas com fim na acessibilidade do conhecimento construído a partir da pesquisa científica.

Ademais, as questões e os objetivos estabelecidos para essa pesquisa vão ao encontro das percepções de Borgman, Scharnhorst e Golshan (2019), onde arquivos/repositórios de dados digitais executam papéis centrais nas infraestruturas do conhecimento como entidades que facilitam o fluxo de dados entre as partes, geralmente ao longo do tempo. Apesar do crescimento de pesquisas sobre práticas, compartilhamento, e reúso de dados, e dos avanços em padrões e normas por meio de organizações como a *Research Data Alliance* (RDA) e a *Force11* [Princípios FAIR], poucos têm estudado o papel dos arquivos/repositórios de dados nas infraestruturas do conhecimento (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 888-889, tradução nossa).

É a partir da afirmação de Borgman, Scharnhorst e Golshan (2019) que se justifica a investigação proposta, da percepção das necessidades de exploração dos assuntos e da procura por soluções criativas à abordagem dos problemas que envolvem os dados científicos.

1.4 Estrutura da dissertação

O Capítulo 2 aborda conceitualmente alguns pontos-chave da pesquisa desenvolvida, passando por discussões quanto às definições do que são dados, informação e conhecimento. Também se discute o fenômeno *Big Data* e suas implicações na sociedade moderna; e a Ciência de Dados (*Data Science*) enquanto área do conhecimento que se entrelaça à CI em pontos específicos. Apresenta-se a *e-Science* ou quarto paradigma da ciência, bem como os serviços voltados aos dados científicos que surgem a partir dele. Há também a apresentação das proposições e diretrizes que formatam os Princípios FAIR. Como passo final, o capítulo reúne e expõe as principais ideias dos autores acessados a partir do levantamento bibliográfico de forma a estabelecer uma estrutura ao alcance dos objetivos propostos.

Em sequência, o Capítulo 3 apresenta a metodologia de pesquisa aplicada ao desenvolvimento da investigação em questão, onde se descrevem os passos tomados no levantamento bibliográfico que serve à revisão, na seleção dos objetos estudados, na coleta dos dados investigados, e na descrição das técnicas utilizadas para sua análise.

Adiante, a responsabilidade pela entrega dos resultados obtidos por meio da análise dos dados da pesquisa está imbuída ao Capítulo 4, apresentada em três etapas.

Ao quinto e último capítulo são reservadas as considerações finais dessa investigação.

2 REFERENCIAL TEÓRICO E REVISÃO DE LITERATURA

Com base em textos oriundos da CI, este capítulo propõe uma construção conceitual, buscando entender os objetos estudados de forma intrínseca, em movimento de externalização desse conhecimento, a partir da percepção dos fenômenos em congruência com o campo da CI, assim como no que compete à Computação e outros campos correlatos.

2.1 Referencial teórico

2.1.1 Big Data

Organizações do conhecimento estão incorporando práticas destinadas ao uso dos dados em seu dia a dia, uma vez que esses são a base da pirâmide do conhecimento. Considerando a Gestão do Conhecimento como Gestão do Conhecimento Tácito ou Cognitivo, Gestão da Informação como Gestão do Conhecimento Documentado, Estruturado e Explícito, então, Gestão de Dados seria, em grande parte, Gestão do Conhecimento Não Estruturado e Explícito. De acordo com o primeiro e mais comum modelo dentro do estudo de Zins (2007), o conhecimento está no Domínio Subjetivo (DS), o qual traz entendimento que se trata de um fenômeno intrínseco, não podendo ser resultado de dados por si só, embora possa ser criado por meio da análise e interpretação dos dados (internalização), transformando-os em capital intelectual das organizações (ZINS, 2007, p. 489, tradução nossa).

A pergunta a se fazer é “o que são dados?”. O único acordo nas definições é que nenhuma definição será suficiente. Dados têm vários tipos de valor, e esse valor pode não ser aparente até muito depois deles serem coletados, tratados, ou perdidos. O valor dos dados varia muito em relação ao local, hora, e contexto (BORGMAN, 2015, p. 20, tradução nossa).

Em reposta indireta ao questionamento de Borgman (2015), Amaral (2016) afirma em seu livro que “dados são fatos coletados e normalmente armazenados. Informação é o dado analisado e com algum significado. O conhecimento é a informação interpretada, entendida e aplicada para um fim. [...] O dado pode estar em formato eletrônico analógico ou digital”. (AMARAL, 2016, p. 3).

Por dado eletrônico, pode-se compreender todo dado coletado que venha a ser digitalizado, ou seja, registrado mais comumente por anotações em meio físico e transformado em digital (eletrônico analógico), ou dados nato digitais, que nascem em meio digital por coletas feitas por humanos, máquinas, sensores, robôs, entre outros (eletrônico digital). “O dado digital é todo aquele armazenado na forma de ‘zeros e uns’, independente de sua estrutura. [...] informação estruturada em planilha eletrônica é dado. Vídeos digitais,

postagens em redes sociais, dados de acelerômetros em um celular [...]" entre outros (AMARAL, 2016, p. 4).

Ilharco (2004) levanta também alguns questionamentos relevantes no que tange à informação: "[...] a sociedade da informação é a sociedade de quê? O que é a informação? Quais os seus princípios de base? As suas relações com fenômenos próximos, como a comunicação, a ação, o conhecimento, a nova tecnologia?". (ILHARCO, 2004, p. 1). Essas são perguntas relevantes e que também podem ser utilizadas e adaptadas ao contexto dos dados, conforme proposto por Borgman (2015).

No centro do problema da curadoria de dados estão perguntas como: quais dados são dignos de preservação, por quê, para quem, por quem, e por quanto tempo? Quais responsabilidades da curadoria de dados devem recair sobre investigadores, comunidades, universidades, agências financiadoras, ou outros *stakeholders*? (BORGMAN, 2015, p. 29, tradução nossa).

Borgman, Scharnhorst e Golshan (2019) definem os *stakeholders* como sendo acadêmicos e equipes que produzem dados, agências de financiamento que provêm recursos à condução de pesquisas, universidades e outras instituições de pesquisa, produtores de políticas de pesquisa em organizações públicas e privadas, usuários desses dados, bibliotecas e arquivos que podem adquirir e gerir dados (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 888, tradução nossa).

Quando da discussão de fenômenos próximos à informação como indagado por Ilharco (2004), pode-se destacar um em alta atualmente, o *Big Data*. *A*ntes de qualquer definição desse termo, é preciso enfatizar que "*Big Data* é um fenômeno e não tecnologia". (AMARAL, 2016, p. 9). Com isso em mente, pode-se dizer que o volume, velocidade, variedade, veracidade e valor dos dados, os 5 Vs do *Big Data* (inicialmente descritos em 3 Vs: volume, velocidade e variedade) são um obstáculo para aqueles que buscam gerir conhecimento em termos de inteligência competitiva ou de negócio, pois "os dados devem ser obtidos, processados, e efetivamente usados, levantando problemas relacionados a como o *Big Data* será representado e modelado". (STOREY; SONG, 2017, p. 51, tradução nossa).

De acordo com o levantamento bibliográfico realizado por Mashingaidze e Backhouse (2017), a inteligência de negócio é um conjunto de aplicações, tecnologias, arquiteturas, processos e metodologias usadas para coletar, armazenar, recuperar e analisar dados (HU *et al.*, 2014; GUPTA *et al.*, 2015 *apud* MASHINGAIDZE; BACKHOUSE, 2017, p. 493, tradução nossa).

A inteligência competitiva e a Ciência de Dados surgem na transição do milênio como teorias, conceitos e práticas direcionadas principalmente ao lucro das organizações, enquanto que a *e-Science* ou e-Ciência (também conhecida como o quarto paradigma da

ciência ou ciência 2.0) e os serviços de dados científicos (entre eles a gestão e a curadoria) surgem como teorias, conceitos e práticas direcionadas à evolução do fazer científico.

Em muitas instâncias, a ciência está atrasada em relação ao mundo comercial na habilidade de inferir significado a partir dos dados e agir com base nesse significado. A maioria dos dados científicos não possui um valor econômico alto o suficiente para estimular um desenvolvimento mais rápido de descobertas científicas (HEY; TANSLEY; TOLLE, 2009, p. xi, tradução nossa).

Dados são parte de um fenômeno tanto quanto a informação também o é, pois ela se origina a partir da análise e interpretação de dados ou conjuntos de dados em determinado formato e/ou mídia, que por sua vez são compreendidos dentro de um ou vários contextos. Uma definição simplória de *Big Data* seria um grande volume de dados com origem em diversas fontes, e que são estruturados ou não, públicos ou privados. Wamba e outros (2015) definem "[...] '*Big Data'* como uma abordagem holística para gerir, processar e analisar os 5 Vs [...] com o intuito de criar *insights* acionáveis para a entrega de valor sustentado, medindo o desempenho e estabelecendo vantagens competitivas". (WAMBA *et al*., 2015, p. 235, tradução nossa).

De acordo com Storey e Song (2017), *Big Data* se refere a grandes quantidades de dados, os quais organizações são capazes de capturar e analisar de forma significativa para que assim decisões baseadas em dados possam ser tomadas. O volume de dados tem crescido exponencialmente nas últimas décadas, ao ponto em que o gerenciamento desse ativo (dados) por meios tradicionais não seja mais possível (STOREY; SONG, 2017, p. 50, tradução nossa).

Profissionais de áreas distintas (Ciência da Computação, Estatística, Sistemas de Informação, Ciência da Informação, entre outras) estão lidando ou sendo imbuídos com a tarefa desafiadora de gerir esses conjuntos de dados, focando seus esforços na criação de uma cultura organizacional inovadora, ao desenvolver melhores práticas, mudando assim as formas de pensar e resolver problemas com vistas à vantagem competitiva. "O *Big Data* vai oferecer muitas oportunidades. Estas oportunidades virão de duas formas: vantagem competitiva ou criação de produtos e/ou serviços orientados a dados". (AMARAL, 2016, p. 11). Dados são, portanto, ativos a serem explorados.

> [...] hoje dados são produzidos massivamente em redes sociais, comunidades virtuais, blogs, dispositivos médicos, TVs digitais, cartões inteligentes, sensores em carros, trens e aviões, leitores de código de barra e identificadores por radiofrequência, câmeras de vigilância, celulares, sistemas informatizados, satélites, entre outros (AMARAL, 2016, p. 8).

De acordo com Swan (2015), *Big Data* é um enorme conjunto de dados que pode ser grande em volume, velocidade, variedade, veracidade, e variabilidade. Os volumes e

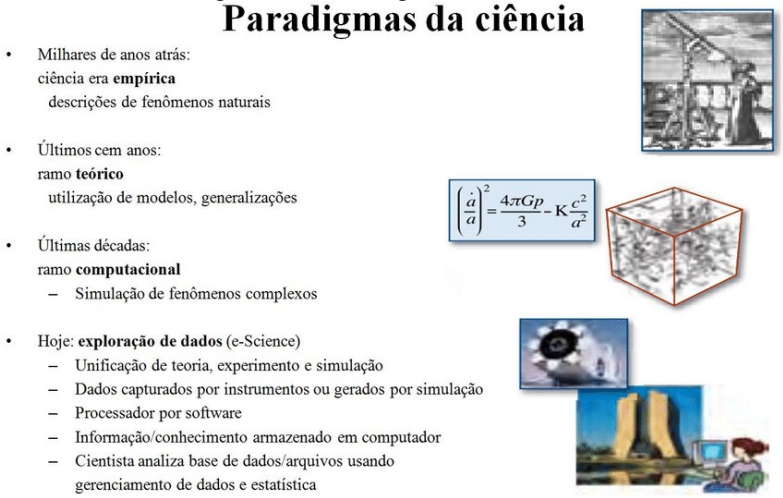
atividades de dados são similarmente “grandes” em quatro áreas: científica, governamental, corporativa, e dados pessoais (SWAN, 2015, p. 1, tradução nossa).

Big Data envolve o uso de diversos tipos de conceitos e tecnologias, como computação nas nuvens, virtualização, *Internet*, estatística, infraestrutura, armazenamento, processamento, governança e gestão de projetos (AMARAL, 2016, p. 9).

2.1.2 e-Ciência e Ciência de Dados

A mais de uma década atrás, em palestra na *Computer Science and Telecommunications Board* (2007), o engenheiro da computação Jim Gray abordava o surgimento de um novo paradigma, a *e-Science*, ou melhor, o quarto paradigma da ciência. A Figura 2 apresenta a evolução dos paradigmas da ciência aos dias atuais, conforme Hey, Tansley e Tolle (2009) em *The Fourth Paradigm: data-intensive scientific discovery*[3].

Figura 2 – Paradigmas da ciência.



Fonte: HEY; TANSLEY; TOLLE, 2009, p. xviii.

Esse novo paradigma se apoia nos anteriores, contudo se utiliza do *Big Data* para que a partir de seu processamento, o pesquisador possa gerar conhecimento por meio de análises feitas em cem por cento do universo pesquisado, e não somente em amostras como praticado por séculos pela ciência tradicional (HEY; TANSLEY; TOLLE, 2009, p. xiii, tradução nossa).

Nossa capacidade de medir, armazenar, analisar, e visualizar dados é a nova realidade para a qual a ciência deve se adaptar. Os dados estão no centro desse novo paradigma, e se sentam ao lado do empirismo, da teoria, e da simulação, que juntos formam

[3] O Quarto Paradigma: descobertas científicas na era da *eScience*. Versão publicada em português pela Oficina de textos, 2011.

o *continuum* considerado como o método científico moderno (HEY; TANSLEY; TOLLE, 2009, p. 210, tradução nossa).

A ciência intensiva ou baseada em dados consiste em três atividades básicas, que são respectivamente a captura, a curadoria, e a análise (HEY; TANSLEY; TOLLE, 2009, p. xiii, tradução nossa). Em paralelo, na visão de Swan (2015), observa-se que a ciência intensiva em dados (também *e-Science*) é uma ciência computacionalmente intensiva envolvendo enormes conjuntos de dados que podem requerer técnicas de computação em Ciência de Dados para modelagem, observação, e experimentação de alta dimensão, e pode ser realizada em ambientes de redes distribuídas (SWAN, 2015, p. 2, tradução nossa).

Organizações dispostas a inovar devem olhar o *Big Data* de forma diferenciada para que dessa maneira possam conhecer oportunidades e ameaças num mundo modernamente digital. A solução que vem com esse desafio se encontra em diversos campos do conhecimento, a qual se torna monetariamente inviável a muitas organizações devido aos custos de contratação de mão de obra especializada. Com isso, as descrições de ofertas de emprego começam a direcionar essas competências multidisciplinares a apenas um único profissional, aquele que seria responsável por responder às demandas geradas por dados. Alguns consideram essa uma tarefa impossível, relacionando esse profissional desejável a seres místicos retirados de folclores.

Dada a dimensão das habilidades requeridas, conclui-se que possa não ser realista esperar que qualquer pessoa possua todas as especialidades relevantes. Dado seu quase *status* místico, um número crescente de profissionais de dados está começando a se referir a tais raros indivíduos, que dizem se destacar em uma ampla gama de disciplinas tradicionalmente distintas, como unicórnios (BASKARADA; KORONIOS, 2016, p. 65, tradução nossa).

De acordo com os resultados de pesquisa de Baskarada e Koronios (2016), todos entrevistados concordaram que é improvável esperar que uma pessoa tenha o mesmo nível de *expertise* em um número distinto de disciplinas por mais especializados que sejam. Ao invés disso, todos eles buscaram construir efetivos times multidisciplinares (BASKARADA; KORONIOS, 2016, p. 67, tradução nossa). Essas equipes consistiriam em profissionais advindos de diversos campos de atuação, possuindo competências definidas como habilidades no trabalho de Baskarada e Koronios (2016) apresentadas no Quadro 1.

Independentemente desses profissionais existirem ou não, as organizações do conhecimento os classificam como cientistas de dados, termo provavelmente cunhado na academia, baseando-se em Ciência de Dados enquanto campo de pesquisa, mas que se expandiu a outros ambientes tecnológicos.

Para Patil e Davenport (2012), a Ciência de Dados é uma disciplina aplicada emergente que busca facilitar tomadas de decisão organizacional por meio do

desenvolvimento de modelos estatísticos que extraem conhecimento de dados brutos (PATIL; DAVENPORT, 2012 *apud* BASKARADA; KORONIOS, 2016, p. 65, tradução nossa). Na concepção de Amaral (2016), pode-se "definir Ciência de Dados como os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida: da produção ao descarte". (AMARAL, 2016, p. 6).

Por conseguinte, habilidades requeridas que se encaixam na aplicação da Ciência de Dados (considerando sua essência multidisciplinar) se apoiariam em Matemática e Estatística (ex.: mineração de dados, teste de hipóteses, e análise preditiva), Ciência da Computação (ex.: estruturas e algoritmos de dados), e *expertise* de domínio. Outras habilidades podem possibilitar a integração, transformação, e carregamento de dados, assim como a visualização de dados (BASKARADA; KORONIOS, 2016, p. 66, tradução nossa).

Quadro 1 – Funções e habilidades em equipes de Ciência de Dados.

Conjuntos de habilidades	**Especialista de domínio**	**Engenheiro de dados**	**Estatístico**	**Cientista da computação**	**Comunicador**	**Líder da equipe**
Habilidade(s) n.1	Entendimento do processo	Extração de dados	Refinar e formalizar questões e ideias de especialistas de domínio	Proficiência em ferramentas e tecnologias: ex.: *Apache Hadoop, Map reduce*, e *Spark*	Exploração de problemas e oportunidades relevantes	Entendimento de todas as outras funções em razão de reunir a equipe, gerir recursos, tarefas, e entregáveis
Habilidade(s) n.2	Entendimento de política governamental	Limpeza de dados	Requisitar dados relevantes de engenheiros de dados	Linguagens de programação relevantes como R e *Python*	Comunicação de eventuais resultados	*Expertise* extensa em gestão de projetos
Habilidade(s) n.3	Fazer perguntas a questões relevantes	Enriqueciment o de dados	Guiar cientistas da computação em relação à análise dos dados	Computação em *cluster* e na nuvem	Contar histórias (*Storytelling*)	Responsável por assegurar que qualquer norma de ética, de privacidade, de segurança, e expectativas sejam aderidas
Habilidade(s) n.4	Gerar hipóteses relevantes	Transformação de dados	*Design* experimental	Processamento (ex.: ordenar, agregar, buscar, combinar e concatenar)		
Habilidade(s) n.5	Interpretar resultados		Teste de hipóteses	Análise de enormes conjuntos de dados		

Fonte: Baskarada; Koronios, 2016, p. 68-69, tradução nossa.

Para além das funções e habilidades compiladas no quadro demonstrado, há outras que também merecem destaque, pois se encontram dentro do processo da preservação digital e ciência aberta, cujas práticas se tornam igualmente importantes à Ciência de Dados. Tais práticas envolvem organização, preservação e compartilhamento de dados, informação e conhecimento, assim como competências em pesquisa científica e conhecimento de domínio.

Os interesses dos cientistas de dados – cientistas da computação e da informação, engenheiros de *software* e de base de dados, programadores, especialistas de domínio, curadores e especialistas em anotações, bibliotecários, arquivistas, e outros cruciais à gestão bem sucedida de coleções de dados digitais – estão na obtenção do reconhecimento pleno de suas contribuições intelectuais e de sua criatividade (NCAR, 20--? apud HEY; TANSLEY; TOLLE, 2009, p. xii, tradução nossa).

Há que se abrir parênteses para uma observação de Amaral (2016), num contraponto a uma parte da literatura de áreas como a Ciência da Computação, Sistemas da Informação e afins.

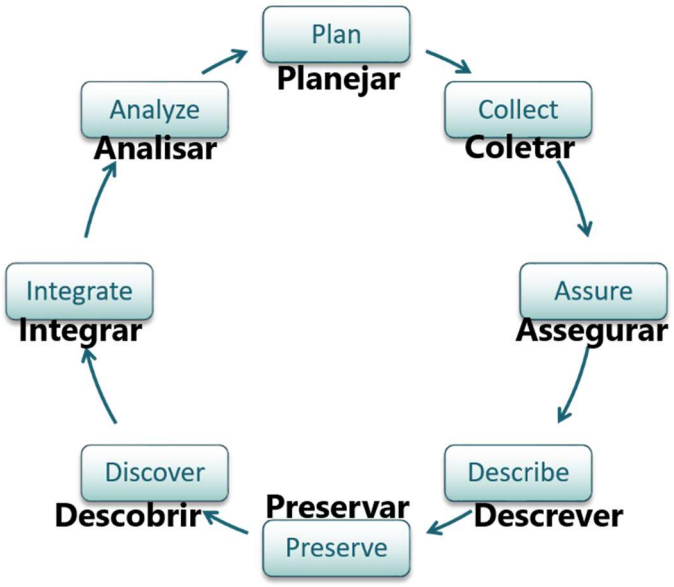
> Normalmente, a Ciência de Dados é associada de forma equivocada apenas aos processos de análise dos dados, onde com o uso de estatística, aprendizado de máquina ou a simples aplicação de um filtro se produz informação e conhecimento. Nessa visão "míope", a Ciência de Dados passa a ser vista apenas como um nome mais elegante para Estatística. Antes de tentarmos entender o porquê da Ciência de Dados não ser a mesma coisa que Estatística, precisamos compreender o ciclo de vida do dado (AMARAL, 2016, p. 4).

Cada etapa no ciclo de vida dos dados vai necessitar da aplicação de técnicas e pessoal específico para gerenciar e garantir que elas sejam cumpridas, até que se alcance o objetivo proposto. "Há uma série de concepções de modelos de ciclo de vida de dados de pesquisa, cada um com particularidades e objetivos determinados, muitas vezes orientados para domínios de conhecimento específicos". (SAYÃO; SALES, 2015, p. 11). "Dentro do ciclo de vida dos dados existe um começo, meio, fim e recomeço. De forma simplória, pode-se dizer que os dados são produzidos ou coletados, armazenados, transformados, analisados, visualizados, e por fim, descartados". (AMARAL, 2016, p. 5-6).

Concernindo a proteção de ativos intangíveis em instituições do conhecimento, Amaral (2016) coloca que "[...] o dado, enquanto existente, terá a ele associado questões de segurança, privacidade e qualidade. Ainda, dados dentro de uma organização são governados por políticas e procedimentos, mesmo que informais". (AMARAL, 2016, p. 5). Uma política institucional voltada à Ciência de Dados deve se preocupar com o ciclo de vida dos dados, que como percebido, não se resume à coleta e análise. Desse modo, pode-se usufruir da máxima capacidade dos dados, seja por uso ou reúso.

A Figura 3 nos apresenta um modelo de ciclo de vida dos dados criado pelo *Data Observation Network* (*DataONE*). Cada modelo poderá ser adaptado de acordo com as necessidades de comunidades em diferentes domínios.

Figura 3 – Ciclo de vida dos dados de acordo com o *DataONE*.



Fonte: *DataONE*, 2012, tradução nossa.

O *DataONE* é uma rede composta por parcerias estabelecidas entre organizações participantes que possuem *expertise* de muitas décadas em uma ampla variedade de campos como iniciativas de arquivos, bibliotecas, sistemas de observação ambiental e redes de pesquisa, gestão de dados e informação, centros de síntese científica, e sociedades profissionais (*DATAONE*, 2012, *online*, tradução nossa). Nela, observamos diferenças entre o modelo proposto por Amaral (2016), pois aqui percebemos um ciclo em domínio específico (Biologia e Ciências da Natureza) se iniciando em passos expressos por ações que buscam: planejar, coletar, assegurar, descrever, preservar, descobrir, integrar, e analisar os dados.

2.1.3 Serviços de dados científicos

Devido ao dilúvio de dados no século XXI, profissionais precisaram desviar sua atenção dos serviços dirigidos a informações para aqueles direcionados à manipulação de dados. Por conta disso, tais serviços precisaram passar por remodelação para que dessa maneira pudessem responder às necessidades de dados, provocando o surgimento de novas práticas profissionais, o que resulta em novos papéis e carreiras.

Há uma preocupação relacionada a como profissionais da informação estão se comportando frente a esses desafios. Discutir tópicos relacionados a dados se torna relevante no intuito de incitar ação e conscientização profissional.

Dados científicos são resultado do fazer científico, considerando que essa prática se utiliza da computação como meio para um fim, fazendo computadores processarem simulações, gerando *petabytes* de dados no mundo inteiro todos os dias. A ciência, como se sabe, não apenas simula, mas também registra experimentos em vários campos de pesquisa,

mirando sempre a criação de novos conhecimentos baseados em análise e interpretação dos dados.

A gestão e curadoria de dados científicos envolvem serviços (atividades no Quadro 2), ferramentas e infraestruturas do conhecimento que abrangem o ciclo de vida da pesquisa científica. Pesquisadores necessitam de apoio nos processos de planejamento, gestão, organização, documentação e preservação de seus conjuntos de dados, bem como em questões relacionadas a licenças e propriedade intelectual, quando na intenção do compartilhamento de seus dados.

Quadro 2 – Serviços de dados científicos e suas atividades.

Atividades SDC	**Gestão de Dados Científicos**	**Curadoria de Dados**	**Administração de dados**	**Educação da competência em dados**	**Governança de dados**	**Gestão da Qualidade de dados**
Atividade n.1	Cuidar de dados científicos	Possibilitar reúso de dados	Preservação da integridade dos dados	Empoderar indivíduos a transformar dados em informação e em conhecimento acionável	Providenciar dados abertos	Avaliar fontes de dados para fidedignidade e para erros ou problemas
Atividade n.2	Facilitar acesso a dados	Planejar dados	Preservação do acesso a dados	Possibilitar indivíduos a acessar dados	Possibilitar melhor tomada de decisão	Uso de ontologias
Atividade n.3	Preservar dados	Adquirir dados	Dar atenção a quais dados são salvos por seus criadores	Possibilitar indivíduos a interpretar dados	Proteger as necessidades dos *stakeholders*	Organizar dados
Atividade n.4	Adicionar valor a dados	Preparar dados	Dar atenção aonde dados são armazenados e preservados	Possibilitar indivíduos a criticamente avaliar, gerir e usar dados eticamente	Providenciar respostas a questões quanto à disponibilidade e possibilidades de acesso, procedência, fidedignidade e significado	Marcar dados [t*agging*]
Atividade n.5	Hospedar dados	Analisar dados	Dar atenção a como dados são descritos, descobertos, acessados e reusados	Avaliar a qualidade dos dados	Prever o uso indevido de conjuntos de dados institucionais	Rastrear dados
Atividade n.6	Selecionar repositório	Preservar dados			Encorajar o uso efetivo de ativos de dados por instituições	Minerar dados para descobrir padrões em enormes conjuntos de dados
Atividade n.7		Descobrir dados			Providenciar metadados	Visualizar dados

Fonte: Koltay, 2017, p. 347-348, tradução nossa.

Como apresentado no Quadro 2, a gestão e curadoria de dados científicos são serviços fundamentais à organização, preservação, recuperação e reúso desses dados. A gestão e curadoria trazem consigo trabalho intelectual e abordagem técnica no manuseio de dados. A gestão de dados está para a organização, assim como a curadoria está para a preservação em longo prazo, porém não limitada a ela. Padrões de metadados de

preservação digital (como o PREMIS[4]) existem para garantir uma mínima qualidade de descrição focada na preservação em longo prazo.

A curadoria cobre uma ampla gama de atividades, a começar por encontrar as corretas estruturas de dados para mapear em vários armazenamentos. Ela inclui os esquemas e os metadados necessários à longevidade e integração entre instrumentos, experimentos e laboratórios (HEY; TANSLEY; TOLLE, 2009, p. xiii, tradução nossa).

Destaca-se o termo "competência em dados" (*data literacy*) apresentado por Koltay (KOLTAY, 2015, p. 401, tradução nossa). Esse termo é conhecido de áreas como a Ciência da Computação, Sistemas de Informação, entre outras. No entanto, nesse ambiente computacional, *data literacy* denota a formação e o desenvolvimento (letramento) de habilidades técnicas em análises quantitativas (matemáticas e estatísticas) de dados, e a apresentação de resultados de formas mais inteligíveis à compreensão humana. Na CI se trabalha outro termo, "competência informacional", que não pode ser adaptado à "competência em dados", ainda que se reconheça que dado gere informação, e informação gere conhecimento.

O campo da competência informacional pode não concordar com a definição precisa de competência informacional, mas a maioria das pessoas usa o termo "competência informacional" ao invés de "instrução da biblioteca[5]" ou "fluência informacional". O que tem sido chamado "competência em dados" envolve competência estatística, raciocínio quantitativo ou competência quantitativa, e *numeracy*[6]*,* todos significando a mesma coisa (HUNT, 2004, p. 14, tradução nossa).

O termo "competência em dados" na Ciência da Informação necessita ser definido a partir de conceitos e técnicas praticadas por profissionais com *expertise* em organização e preservação de dados digitais (científicos, governamentais, mercadológicos, financeiros, etc.). Isso significa conhecer e mapear o fluxo de trabalho científico e necessidades da comunidade alvo; depositar e descrever dados de forma padronizada com uso de vocabulários controlados e padrões de metadados universais e de domínio; atribuir identificadores persistentes a objetos digitais (conjuntos de dados); vincular publicações e dados a outros relacionados; conhecer leis de proteção a dados e propriedade intelectual; dominar ferramentas e *software* de preservação digital; entre outros. Não se objetiva esboçar uma definição para o termo *data literacy* no contexto da CI.

---

[4] Padrão internacional de metadados que apoia a preservação de objetos digitais e assegura sua usabilidade em longo prazo (*LIBRARY OF CONGRESS*, 2020, *online*, tradução nossa).

[5] O termo original utilizado pela autora é *library instruction*, que também poderia ser traduzido como "instrução bibliotecária", porém se confundiria com o profissional dessa área.

[6] Como o termo *literacy* significa letramento, o termo *numeracy* denota competência com números e matemática.

Sabe-se que repositórios de documentos bibliográficos digitais no Brasil são uma realidade, enquanto repositórios de dados científicos se encontram em processo de estudo e discussão, embora muito ainda tenha que ser estabelecido. Profissionais que buscarem trabalhar com dados encontrarão solo fértil em laboratórios de pesquisa, agências de fomento, unidades de informação e em organizações do conhecimento, assim como em equipes heterogêneas de Ciência de Dados, passando dessa maneira a encarar o desafio de possuir habilidades computacionais aliadas à *expertise* de domínio.

Bibliotecários e profissionais da Ciência da Informação podem contribuir de forma vital com a curadoria de dados, a preservação, e com habilidades em arquivamento para garantir custódia segura da produção de pesquisa. Eles podem também providenciar suporte ao engajamento público com ciência, assim como facilitar acesso público a conjuntos de dados científicos (BASKARADA; KORONIOS, 2016, p. 66, tradução nossa).

2.1.4 Os Princípios FAIR

Inicialmente, precisa-se compreender o termo inglês *fair*, enquanto adjetivo, que pode ser traduzido em português, qualificando um substantivo por justo, leal, bom, honesto, sério, entre outros. Na literatura da área, outro termo muito utilizado é *fairness*, quando pesquisadores se referem aos Princípios FAIR como entidade, transformando o adjetivo *fair* em substantivo abstrato, podendo ser traduzido por equidade, justiça, etc. Para a presente pesquisa adotar-se-á a expressão “nível FAIR” ao tratar de *fairness*, evitando assim o aportuguesamento do termo.

Para garantir um satisfatório nível FAIR dos (meta)dados é preciso que esses registros sejam: Encontráveis (*Findable*), Acessíveis (*Accessible*), Interoperáveis (*Interoperable*), e Reusáveis (*Reusable*).

Em 2016, os “Princípios FAIR para a gestão de dados científicos e governança” foram publicados na *Scientific Data*. Os autores pretendiam proporcionar orientações para melhorar a descoberta, acessibilidade, interoperabilidade e reúso de ativos digitais. Os princípios enfatizam a possibilidade da ação-por-máquina (ex.: a capacidade dos sistemas computacionais para encontrar, acessar, interoperar, e reusar dados com nenhuma ou mínima intervenção humana), pois humanos dependem cada vez mais de suporte computacional para lidar com dados como resultado do aumento do volume, complexidade, e velocidade de criação de dados (*GO FAIR*, 2019, tradução nossa).

A fim de descobrir dados relevantes, executar análise de máquina em escala ou empregar técnicas como inteligência artificial para identificar padrões e correlações não visíveis aos olhos humanos, necessita-se de dados bem descritos e acessíveis que estejam em conformidade com os padrões de suas respectivas comunidades. Os Princípios FAIR

articulam os atributos que os dados precisam ter para permitir e aprimorar seu reúso, por humanos e máquinas. Há necessidade de várias coisas, incluindo informações contextuais e de apoio (metadados) para permitir que esses dados sejam descobertos, compreendidos e usados (*EUROPEAN COMMISSION*, 2018, p. 18, tradução nossa).

Dessa forma, os Princípios FAIR são descritos e definidos conforme texto encontrado no *Website* da *Go FAIR*, apresentados no Quadro 3.

Quadro 3 – Princípios FAIR.

Findable **(Encontráveis)**	O primeiro passo em re(usar) dados é encontrá-los. Metadados e dados deveriam ser fáceis de encontrar tanto para humanos quanto para computadores. Metadados legíveis por máquina são essenciais a descobertas automáticas de conjuntos de dados e serviços, então esse é um componente essencial do processo de adaptação FAIR[7]. F.1 Aos (meta)dados são atribuídos um identificador único e persistente globalmente; F.2 Dados são descritos com metadados valiosos (definidos por R1 abaixo); F.3 Metadados incluem claramente e explicitamente o identificador dos dados que eles descrevem; F.4 Metadados são registrados e indexados em um recurso pesquisável.
Accessible **(Acessíveis)**	Uma vez que o usuário encontra os dados necessários, ele(a) precisa saber como eles podem ser acessados, possivelmente incluindo autenticação e autorização. A.1 (Meta)dados são recuperáveis por meio de seus identificadores usando um protocolo de comunicação padronizado; A.1.1 O protocolo é aberto, gratuito e universalmente implementável; A.1.2 O protocolo permite um procedimento de autenticação e autorização, quando necessário; A.2 Metadados são acessíveis, mesmo quando os dados não estão mais disponíveis.
Interoperable **(Interoperáveis)**	Os dados usualmente precisam estar integrados com outros dados. Em adição, os dados precisam interoperar com aplicações ou fluxos de trabalho para análise, armazenamento, e processamento. I.1 (Meta)dados usam uma linguagem formal, acessível, compartilhada, e amplamente aplicável para representação do conhecimento; I.2 (Meta)dados usam vocabulários que seguem os Princípios FAIR; I.3 (Meta)dados incluem referências qualificadas a outros (meta)dados.
Reusable **(Reusáveis)**	O último e maior objetivo dos Princípios FAIR é aperfeiçoar o reúso dos dados. Para alcançá-lo, metadados e dados devem ser bem descritos para que assim eles possam ser replicados e/ou combinados em diferentes cenários. R.1 (Meta)dados são ricamente descritos com uma pluralidade de atributos relevantes e precisos; R.1.1 (Meta)dados são liberados com uma clara e acessível licença de uso dos dados; R.1.2 (Meta)dados são associados com proveniência [origem] detalhada; R.1.3 (Meta)dados atendem aos padrões de comunidade de domínio.

Fonte: *Go FAIR*, 2019, tradução nossa.

Em sua recente publicação intitulada FAIR *Data Maturity Model: specification and guidelines*, o RDA FAIR *Data Maturity Model Working Group* apresenta um modelo de maturidade FAIR que especifica cada um dos subprincípios supracitados a fim de facilitar seu entendimento e orientar profissionais/pesquisadores quanto às demandas a serem atendidas para que se alcance meta(dados) FAIR.

Os princípios de dados FAIR se aplicam a metadados, dados, e infraestrutura de suporte (ex.: motores de busca). A maioria dos requisitos para "encontrabilidade" e acessibilidade pode ser alcançada no nível dos metadados. Interoperabilidade e reúso requerem mais esforços no nível dos dados (*GO FAIR*, 2019, tradução nossa). Esses princípios deveriam ser aplicados também a identificadores, *software* e Planos de Gestão de

[7] Tradução livre ao termo em Inglês "*FAIRification*".

Dados (PGDs) que conjuntamente permitem dados serem FAIR (*EUROPEAN COMMISSION*, 2018, p. 11, tradução nossa).

O conceito de dados FAIR não deve ser confundido com o de dados abertos, embora eles possam se aplicar simultaneamente (conceitos complementares). Dados FAIR podem e devem ser protegidos em alguns casos (ex.: dados referentes a patentes, levando em conta leis de propriedade intelectual e industrial que se divergem entre países, etc.), exigindo assim solicitação de acesso junto ao autor/instituição responsável.

Crosas (2019) afirma que dados FAIR não são equivalentes a dados abertos. Em sua perspectiva, a ideia de alcançar meta(dados) FAIR não passa de um anseio, pois esses nunca são cem por cento FAIR. Coloca que, ao publicar dados restritos, licenças e acordos de uso dos dados devem ser claramente definidos pelos autores ou provedores de dados (CROSAS, 2019, *online*, tradução nossa). Conjuntos de dados FAIR que tiveram arquivos de dados deletados ou perdidos terão seus metadados acessíveis. Idealmente, esses metadados deveriam indicar os motivos da indisponibilidade de seus dados (*RESEARCH DATA ALLIANCE*, 2020, p. 20, tradução nossa).

De acordo com a Comissão Europeia (2018, p. 10), a implementação de dados FAIR precisa andar de mãos dadas com a ideia de que dados criados por pesquisa financiada com verba pública devem ser tão abertos quanto possível, e tão protegidos quanto necessário (*EUROPEAN COMMISSION*, 2018, p. 10, tradução nossa).

2.2 Revisão de literatura

A esse subcapítulo se reservam as principais ideias retiradas da literatura consultada de acordo com os critérios estabelecidos ao levantamento bibliográfico (apresentados no capítulo de metodologia científica) e em alinhamento com os objetivos dessa investigação. Os textos são organizados por data de publicação e ordem alfabética de autor(es). O objetivo desse formato de apresentação busca evidenciar o desenvolvimento das principais discussões e assuntos abordados em relação aos dados científicos nos últimos anos.

2.2.1 Doorn e Tjalsma (2007)

De acordo com os autores, o arquivamento de dados científicos é influenciado pelas dinâmicas da tecnologia da informação moderna, e por isso está passando por mudanças velozes. O artigo destaca alguns dos principais problemas, desenvolvimentos e novos métodos nesse campo especial de arquivamento. Doorn e Tjalsma (2007) investigam como e por que arquivos/depósitos de dados científicos se diferem de escritórios de registro público (DOORN; TJALSMA, 2007, p. 1, tradução nossa).

Projetos de retroarquivamento desenvolveram métodos como o *Atempo Digital Archive* (ADA), que possibilita o arquivamento, *backup*, sincronização, migração e cópia de grandes volumes de dados não estruturados (DOORN; TJALSMA, 2007, p. 2, tradução nossa).

Desde o início de seu depósito, dados têm sido descritos de forma padronizada, o que gera problemas em determinadas áreas como as Ciências Sociais, pois informações sistemáticas sobre a coleta de dados nesse campo têm vital relevância. A descrição dos dados sofreu grande influência da tecnologia, o que levou ao desenvolvimento de esquemas orientados a arquivos de dados como o *Study Description Scheme* (SDS), *Data Documentation Initiative* (DDI) *Metadata System* e também de *software* como *Nesstar* (sistema de recuperação da informação e análise estatística). (DOORN; TJALSMA, 2007, p. 2-7, tradução nossa).

Os dados pessoais envolvem aspectos relacionados à privacidade, que está intimamente relacionada também à questão de proteção e preservação de dados. Uma legislação rigorosa pode significar que o uso de dados pessoais será muito restrito. Esse é um problema complexo e que possui elementos contraditórios. Há uma linha tênue entre proteção de privacidade e preservação de dados pessoais (DOORN; TJALSMA, 2007, p. 3, tradução nossa).

Os autores relatam que os primeiros arquivos de dados foram estabelecidos no início de 1960 e continham dados de *surveys* do campo das Ciências Sociais. Elmo Roper foi fundador da pesquisa de *survey* e responsável por desenvolver depósitos de dados por meio do *Roper Center*. A maioria dos depósitos de dados, cedo ou tarde, afiliou-se a academias ou organizações nacionais de pesquisa (DOORN; TJALSMA, 2007, p. 3, tradução nossa).

Afirmam que a necessidade de criar depósitos surgiu de dados computadorizados sendo utilizados em pesquisas nas Ciências Sociais. As principais razões para preservar dados computadorizados são: permitir verificação dos resultados, uso secundário (reúso) e valor histórico. Vários arquivos surgiram no meio acadêmico (campos como história, literatura e linguística, ciências sociais e humanas, arqueologia, etc.) como depósitos de dados científicos e foram os primeiros institutos a lidar com material eletrônico, além de não terem qualquer conexão com escritórios de registro público (DOORN; TJALSMA, 2007, p. 3-4, tradução nossa).

Nas últimas décadas, a tecnologia computacional tem nos apresentado um mundo cada vez mais digital, de possibilidades imprevisíveis. A recuperação da informação e o armazenamento têm mudado de tal forma que o efeito pode ser comparado ao efeito que a introdução da imprensa teve séculos atrás (DOORN; TJALSMA, 2007, p. 4, tradução nossa).

Há certos obstáculos a serem encarados por profissionais que buscam fornecer acesso permanente a dados digitais como: rápida obsolescência de *hardware* e *software*,

metadados, gestão, propriedade, autenticidade, proteção de privacidade e avaliação. Para os autores, a preservação digital não faz parte do cotidiano de preocupações de profissionais da TI, logo, a solução desses problemas deve vir de arquivistas e bibliotecários. É preciso que esses profissionais tomem a frente e se acostumem com tecnologias da informação (DOORN; TJALSMA, 2007, p. 5, tradução nossa).

Em ambientes acadêmicos ou de pesquisa, a preocupação com preservação digital é maior (caso das bibliotecas universitárias). Investimentos em coleta de dados e digitalização são perdidos caso esses bancos se tornem negligenciados com os anos. As preocupações dos pesquisadores estão na segurança da propriedade intelectual, no tempo de vida de suas publicações eletrônicas e nos dados de sua pesquisa (DOORN; TJALSMA, 2007, p. 5, tradução nossa).

Na preservação digital em longo prazo, existem algumas estratégias para a manutenção do acesso aos materiais: emulação, conversão ou migração, compatibilidade reversa, preservação de *hardware*, identificadores persistentes, impressão de documentos digitais, arqueologia digital, etc. Na estratégia de conversão, os registros são exportados para programas atualizados ou são salvos em formatos abertos (ex.: ASCII, UNICODE, XML[8]). (DOORN; TJALSMA, 2007, p. 6, tradução nossa).

Doorn e Tjalsma (2007) indicam que para acadêmicos, a forma na qual dados são guardados tem relevância secundária, o que importa é ter dados originais. Ademais, é essencial preservar não somente o conteúdo, mas também a forma, manter-se mais próximo do original eletrônico. Conteúdo, estrutura e contexto são atributos de registros eletrônicos (DOORN; TJALSMA, 2007, p. 7-8, tradução nossa).

Um dos problemas na preservação digital: diferentes *stakeholders* durante o ciclo de vida dos documentos eletrônicos. Os criadores dos dados não se sentem responsáveis pelos dados após seu período de uso. Agências de fomento financiam a criação dos dados, e portanto, estão em posição para influenciar políticas diretivas à manutenção deles. (DOORN; TJALSMA, 2007, p. 9, tradução nossa).

O *Open Archival Information System (OAIS)* é um modelo-referência que possui um *background* tecnológico (originou-se da NASA) e pode ser usado no arquivamento de vários materiais desde publicações eletrônicas a dados (sem muitas adaptações). O quadro do OAIS tem se expandido de comunidades de cientistas sociais a grupos de usuários que necessitam assistência na interpretação de dados (DOORN; TJALSMA, 2007, p. 2-9, tradução nossa).

O desenvolvimento de novos esquemas de metadados foram orientados por materiais digitais e a *Internet*, como por exemplo, o *Dublin Core* que originalmente se

[8] *Extensible Markup Language* ou Linguagem de Marcação Extensível.

destinava à descrição de publicações eletrônicas na *Web*, embora possa ser utilizado para descrever objetos e arquivos digitais. Para propósitos de pesquisa científica, esses metadados não fornecem informações detalhadas quanto ao conteúdo dos dados. Outro padrão de metadados, *Metadata Enconding and Transmission Standard* (METS), foi projetado para documentar e organizar materiais digitalizados e espalhados em arquivos diferentes (DOORN; TJALSMA, 2007, p. 10, tradução nossa).

Pelos autores, é inútil manter depósitos de dados científicos se a academia não vê necessidade para tal. Depósitos e suas equipes estão cada vez mais cientes da importância de seu envolvimento tanto no estágio de criação desses arquivos de dados eletrônicos, quanto durante toda sua vida útil (DOORN; TJALSMA, 2007, p. 10-11, tradução nossa).

A infraestrutura dos dados se refere mais do que somente ao arquivamento dos dados. Ela inclui cuidado com os dados desde o momento de sua criação adiante. Doorn e Tjalsma (2007) citam o *Digital Curation Centre* (DCC) como responsável pela primeira definição do termo "curadoria de dados", onde se expressa que "a curadoria digital é sobre preservar e adicionar valor a um corpo de informação digital com vistas em seu presente e futuro uso". (*Digital Curation Centre apud* DOORN; TJALSMA, 2007, p. 13, tradução nossa).

Afirmam que o primeiro propósito dos arquivos digitais científicos e repositórios não é preservar materiais, mas dar acesso ao que é preservado. Citam o crescimento do entendimento comum que o acesso a resultados de pesquisa deve ser mais aberto o possível, levando em conta questões legais de direito à propriedade intelectual e privacidade (DOORN; TJALSMA, 2007, p. 14, tradução nossa).

Embora o acesso aberto lute pelo acesso gratuito sem barreiras, não significa que todas publicações e dados científicos estarão disponíveis. Novas licenças têm sido desenvolvidas para oferecer alternativas ao direito autoral completo. Por exemplo, as licenças *Creative Commons* fornecem uma variedade flexível de proteções e liberdades a autores, artistas e educadores (*Creative Commons apud* DOORN; TJALSMA, 2007, p. 14, tradução nossa).

O objetivo primário da *Open Archives Initiative Protocol for Metadata Harveting*[9] (OAI-PMH) é facilitar o acesso a arquivos eletrônicos em repositórios institucionais interoperáveis para compartilhar, publicar e arquivar metadados. O protocolo OAI surgiu no mundo das impressões eletrônicas (*e-print*), onde o termo "arquivo" é utilizado para indicar um repositório de publicações eletrônicas (embora possa também conter dados, imagens, vídeo, etc.). Metadados de qualquer sistema podem ser incorporados, entretanto para

[9] Mecanismo para a interoperabilidade entre repositórios (*OPEN ARCHIVES*, 2020, *online*, tradução nossa).

alcançar um nível básico de interoperabilidade, é necessário especificar o menor denominador comum, para qual um esquema XML de elementos de metadados *Dublin Core* é utilizado (DOORN; TJALSMA, 2007, p. 15, tradução nossa).

Enquanto dados são cada vez mais armazenados em lugares diferentes, é importante assegurar que eles atendam a mínimos padrões de qualidade, rastreabilidade, acessibilidade e usabilidade. Várias organizações internacionais possuem projetos que procuram especificar critérios para “repositórios digitais confiáveis” (DOORN; TJALSMA, 2007, p. 16, tradução nossa).

2.2.2 Brownlee (2009)

O artigo discute a gestão de metadados, sustentabilidade e níveis de serviços de repositórios. O autor afirma que o número de coleções de dados criados e geridos em unidades acadêmicas de universidades é desconhecido, contudo se assume ser enorme, assim como a variação dos formatos, formulários e ferramentas usadas para capturar, gerenciar, renderizar e manipular os dados (BROWNLEE, 2009, p. 2, tradução nossa).

O autor buscou explorar problemas que concerniam a gestão de dados científicos no repositório de sua biblioteca. O objetivo foi desenvolver guias de forma a assistir uma abordagem sustentável e consistente para lidar com solicitações de gestão de diferentes tipos de materiais (BROWNLEE, 2009, p. 2, tradução nossa).

A primeira parte do trabalho descreve como os dados são criados/coletados e geridos dentro de cada campo de pesquisa e respectivos laboratórios em sua universidade. Em alguns casos, metadados incluem taxonomia desenvolvida pelo próprio pesquisador, em outros, há descrição com o uso de metadados de preservação recomendado pelo Padrão de Metadados da Biblioteca Nacional da Nova Zelândia.

De acordo com o autor, o *DSpace* oferece o *Dublin Core* (DC) como esquema de metadados descritivos pré-definido. Afirma que o DC é adequado à descrição bibliográfica da maioria dos itens, nos casos em que a coleção compreende formatos de publicação tradicional como artigos de pesquisa e anais de eventos (BROWNLEE, 2009, p. 4, tradução nossa).

Na segunda parte, o autor apresenta 4 opções de importação de metadados originais para o *DSpace*, Em cada opção, apontam-se as vantagens e desvantagens de cada prática. Elas são: a) mapear metadados nativos a elementos DC existentes; b) mapear metadados nativos a elementos do DC e criar novos qualificadores customizados para etiquetas DC padrão; c) criar um esquema customizado idêntico ao conjunto de metadados nativo; e d) gerar registros DC como abstrações dos registros de metadados nativos e submeter os originais como objetos digitais suplementares (BROWNLEE, 2009, p. 6, tradução nossa).

Segundo o autor, a primeira opção é a menos provável a satisfazer requisitos para preservação e reúso de metadados de dados científicos. A opção 3 seria a mais cara, embora possa assistir o melhor nível de interatividade e flexibilidade em relação à apresentação. Por fim, a última poderia permitir a menor flexibilidade considerando a interatividade do usuário, apesar de haver incertezas quanto a isso (BROWNLEE, 2009, p. 6, tradução nossa).

O autor em suas considerações finais relata que sua experiência em trabalhar com acadêmicos nas atividades de gestão de dados destacou a necessidade de serviços que deem suporte à *e-Research*. Além dos requisitos de domínio específico para o interrogatório de dados personalizado, ferramentas de apresentação e manipulação, pode haver necessidade comum por serviços que permitam submissão e descrição estruturada e definida por usuários de coleções de dados científicos (BROWNLEE, 2009, p. 8, tradução nossa).

Além da tecnologia e da política, o autor acredita que a gestão de dados requer conhecimento de dados e padrões de documentação associados a um tipo de cientista de dados descrito por Alma Swan em sua classificação. Em sua visão, bibliotecários acadêmicos poderiam se tornar cientistas de dados em seus campos de atuação e pesquisa, podendo desempenhar papel-chave na gestão institucional de dados de domínio específico e em serviços de preservação digital (BROWNLEE, 2009, p. 9, tradução nossa).

2.2.3 Force e Auld (2014)

Trata-se de um breve relato de experiência, onde os autores (empregados da Thomson Reuters) apresentam seu *Data Citation Index*[10], desenvolvido para fornecer a usuários da *Web of Science* uma ferramenta para pesquisar e descobrir dados científicos associados com a pesquisa publicada de maior influência. Conjuntos de dados indexados são vinculados a citações da literatura bibliográfica na *Web of Science*, criando uma visão mais holística do processo de pesquisa acadêmica. O objetivo primário da ferramenta é providenciar informação métrica sobre reúso dos dados (FORCE; AULD, 2014, p. 97, tradução nossa).

De acordo com os autores, atividades voltadas aos dados aumentaram a partir da solicitação de planos de gestão de dados por parte das instituições e financiadores. Eles atribuem o crescimento das discussões sobre dados científicos a pesquisadores preocupados com a descoberta, a atribuição e a citabilidade acadêmica (FORCE; AULD, 2014, p. 97, tradução nossa).

Para que a *Reuters* inclua repositórios de dados em seu *Data Citation Index*, esses precisam atender a certos critérios, como demonstrar estabilidade dos objetos de dados e do repositório que supervisiona sua curadoria, bem como padrões de curadoria e publicação dos

[10] Índice de Citação de Dados.

dados e *links* estabelecidos à pesquisa acadêmica (FORCE; AULD, 2014, p. 97, tradução nossa).

Indiretamente, os autores indicam que os *links* entre publicações científicas e seus dados permitem determinar se o experimento ou estudo é reproduzível. Eles concluem que repositórios de dados disciplinares geralmente diferem nos elementos de metadados disponibilizados e no nível apropriado de granularidade em relação à citação de dados (FORCE; AULD, 2014, p. 97, tradução nossa).

2.2.4 Lee e Stvilia (2014)

Lee e Stvilia (2014) descrevem que agências de financiamento agora solicitam de seus pesquisadores a submissão de planos para a disseminação e o acesso a seus dados científicos, além de vários periódicos e bases de dados já exigirem a publicação de manuscritos com seus respectivos dados (LEE; STVILIA, 2014, p. 1, tradução nossa).

O objetivo do trabalho foi examinar o uso de sistemas de identificação com dados científicos. De acordo com os autores, ao passo que diferentes comunidades gerenciam dados distintos sobre entidades divergentes, esquemas de identificação são contextuais e adaptados a práticas de dados da comunidade (LEE; STVILIA, 2014, p. 1, tradução nossa).

Em seu artigo, os autores citam a *Thompson Reuters* e seu projeto chamado *Data Citation Index*. Afirmam que para fazer a conexão entre publicações e seus dados proposta pela *Reuters*, é essencial o uso de robustos sistemas de identificação. Eles também reiteram que o uso prático de sistemas de identificação para dados científicos não era estudado sistematicamente na literatura. A revisão de literatura desenvolve uma taxonomia de características de sistemas de identificação que pode ser usada por gerentes e curadores de dados na seleção de identificadores para seus repositórios (LEE; STVILIA, 2014, p. 2, tradução nossa).

Entendem que a definição de identificadores deveria mencionar características de sistemas de identificação, tipos de entidade atribuídos, e propósitos de identificadores. Eles definem um identificador de dados como uma sequência de símbolos desenhada para identificar, citar, anotar, e/ou vincular dados científicos a seus metadados. Diferentes sistemas de identificação podem ser usados para referenciar distintos tipos de entidades (LEE; STVILIA, 2014, p. 3, tradução nossa).

Lee e Stvilia (2014) encontraram 14 identificadores na literatura consultada (70 artigos e 40 outras fontes), eles são:

Archival Resource Key (ARK), identifica dados científicos e dá acesso a seus metadados; *Digital Object Identifier* (DOI), identifica objetos e remete a seus registros de metadados bibliográficos; sistema *Handle*, esquema identificador (similar ao DOI) para

recursos da *Internet*, remete a registros de metadados de objetos; *Persistent Uniform Resource Locator* (PURL), identificador vinculado a registros de metadados de objetos; *Uniform Resource Identifier*[11] (URI), categoria de esquemas de identificação para recursos da *Web*, inclui esquemas URL e *Uniform Resource Name*[12] (URN); *Universally Unique Identifier* (UUID), seu uso se expandiu da construção de *software* à identificação de dados, vinculando-os a seus registros de metadados; *National Center for Biotechnology Information* (NCBI) *Accession Number*, identificador único de domínio atribuído a registros de sequência quando submetidos ao *GenBank*[13]; *Chemical Abstracts Service* (CAS) *Registry Number*, seus registros contêm vários tipos de sequências e substâncias orgânicas e inorgânicas únicas; *Life Science Identifier* (LSID), identificador de domínio que identifica e dá acesso a entidades das ciências da vida; *International Standard Name Identifier* (ISNI), desenvolvido pela *International Organization for Standardization* (ISO), identifica identidades públicas em vários campos de atividade criativa; *Open Researcher and Contributor ID* (ORCID), desenvolvido para desambiguar acadêmicos com mesmo nome e fazer conexões entre pesquisas; *ResearcherID*, criada pela *Thompson Reuters* para solucionar a ambiguidade dos nomes de autores em comunicações acadêmicas; *OpenID*, destinado à autenticação de identidade ao se conectar a um *Website*; e *GeoNameID*, sistema de identificação de dados geográficos (LEE; STVILIA, 2014, p. 4-9, tradução nossa).

Dados científicos são materiais factuais registrados comumente aceitos na comunidade científica como indispensáveis para validar achados. Entretanto, os tipos, os formatos e a especialidade necessária para interpretar e curar dados científicos são contextuais e dependentes de domínio. Identificadores de domínio são desenhados para propósitos e necessidades particulares, porém podem ser menos interoperáveis que identificadores genéricos (LEE; STVILIA, 2014, p. 9-10, tradução nossa).

O *Open Archival Information System* (OAIS) é um modelo de referência conceitual da ISO planejado para informar o desenvolvimento de sistemas destinados à curadoria de dados digitais em longo prazo. Liderado pelo OAIS, o *Preservation Metadata: Implementation Strategies* (PREMIS) é um vocabulário de metadados de preservação. Seu modelo de dados consiste em 5 entidades de alto nível: entidades intelectuais, objetos, eventos, agentes, e direitos (LEE; STVILIA, 2014, p. 10, tradução nossa).

Entre os vários tipos de entidades apresentadas no artigo, o tempo é reportado como informação essencial a dados científicos, no entanto os autores afirmam que não há menção do uso de identificadores com entidades de tempo na literatura (LEE; STVILIA, 2014, p. 15, tradução nossa).

---

[11] Identificador de Recurso Uniforme.
[12] Nome de Recurso Uniforme.
[13] Banco de dados de sequência genética.

Pelos autores, assunto é um elemento importante em qualquer esquema de metadados bibliográficos. Catalogadores, curadores, e/ou autores atribuem palavras-chave ou frases a recursos, as quais são utilizadas por usuários para descobri-los. Bibliotecas e comunidades acadêmicas usam diferentes tesauros, vocabulários controlados, e ontologias para reunir dados relacionados, desambiguando e reduzindo a variância de vocabulário nos metadados (LEE; STVILIA, 2014, p. 16, tradução nossa).

O estudo identificou 4 tipos de atividades que utilizam identificadores de dados: identificação, citação, vinculação, e anotação de dados científicos (LEE; STVILIA, 2014, p. 17, tradução nossa).

De acordo com os autores, a rede *Dataverse* é uma aplicação de código aberto que provê diretrizes e ferramentas para a citação de dados. Ela especifica o registro global *Handle* como seu sistema de identificação persistente. O DOI também pode ser utilizado como sistema identificador padrão do *Dataverse*. Os autores fazem menção à comunidade *DataONE*, porém não apresentam informações quanto ao uso de sistemas de identificação por ela. Para Lee e Stvilia (2014), se dados científicos em um repositório de dados não estão associados a artigos relevantes, os dados estão escondidos, limitando seu uso e reúso (LEE; STVILIA, 2014, p. 18-19, tradução nossa).

Os autores discutem 7 dimensões de qualidade dos identificadores, elas são: simplicidade, opacidade, verificabilidade, contextualidade, interoperabilidade, acionabilidade, e granularidade (LEE; STVILIA, 2014, p. 21, tradução nossa).

O estudo examinou a prática do uso de identificadores de dados, e seus resultados forneceram entendimento conceitual baseado na análise da literatura. Indicam necessidade de pesquisa futura com análise de dados empíricos (LEE; STVILIA, 2014, p. 25, tradução nossa).

Concluem que todos os identificadores estudados dão suporte a atividades de identificação, vinculação, e podem ser usados como URI (qualquer identificador com sintaxe URL pode ser utilizado na citação de dados). (LEE; STVILIA, 2014, p. 29, tradução nossa). Para eles, seus achados podem informar *stakeholders* quanto a necessidades e requisitos para que um esquema de identificação possa ajudar a identificar, citar, vincular, e anotar dados científicos (LEE; STVILIA, 2014, p. 33, tradução nossa).

2.2.5 Rousidis e outros (2014)

Para Rousidis e outros (2014), dado o grande volume e diversidade de dados científicos, repositórios de pesquisa estão se tornando parte integral do processo de comunicação e de colaboração entre cientistas e grupos de pesquisa. Ainda, problemas relacionados à qualidade dos dados podem impedir o processo de análise, integração e reúso

de conjuntos de dados heterogêneos. De acordo com os autores, há pouca pesquisa sobre problemas de qualidade dos dados e dos metadados usados para descrever e anotar conjuntos de dados nesses repositórios. O uso de metadados completos e precisos é relevante a vários processos, incluindo reúso e compartilhamento de conjuntos de dados entre pesquisadores; aplicação de curadoria digital e estratégias de proveniência dos dados; e análise de conteúdos de repositórios de dados científicos (ROUSIDIS *et al*., 2014, p. 279-280, tradução nossa).

O objetivo da pesquisa foi identificar os problemas de qualidade dos dados associados com os metadados utilizados em um repositório de dados científicos chamado *Dryad*. Seus objetivos são exploratórios: executar uma análise descritiva dos elementos de metadados usados no *Dryad*; e identificar os principais problemas de qualidade dos metadados (ROUSIDIS *et al*., 2014, p. 280, tradução nossa).

Os autores definem o *Dryad* como um repositório de acesso aberto que permite cientistas – das ciências puras e da medicina – armazenarem, buscarem, recuperarem e reusarem dados científicos associados a suas publicações acadêmicas. Dados são depositados como arquivos com identificadores persistentes (DOIs) e metadados (ROUSIDIS *et al*., 2014, p. 281, tradução nossa).

Dentre todos os elementos de metadados levantados no *Dryad*, eles apresentam três deles (*Creator*, *Type*, *Date*), pois esses representam casos típicos onde problemas de qualidade dos dados podem impedir análises quantitativas e qualitativas no *Dryad*, assim como no reúso dos dados arquivados no repositório. Os problemas encontrados quanto ao elemento *Creator* foram: nomes adicionais, autores que foram registrados apenas com o primeiro nome ou com sobrenome em registro diferente; uso de iniciais; línguas diferentes, variações de escritas de nomes entre idiomas; entrada inválida ou não preenchida; pontos e vírgulas; espaçamento; e miscelânia, uso de texto irrelevante. Quanto ao elemento *Date*, se apresentam na: falta de consistência no formato de data, data não existente, falsa ou não preenchida; e falta de padronização do formato de data (ROUSIDIS *et al*., 2014, p. 283, tradução nossa).

O elemento *Type* ou *DC.type* se refere ao tipo de arquivo. Em seu levantamento, Rousidis e outros (2014) encontraram tipos de arquivos que não deveriam aparecer, como *custom*, *blanks*, *none*, *oneyear*, *protocol* e *untilArticleAppears*. Depois da limpeza, eles descobriram que quase 90% dos registros eram de tipo *dataset* (conjuntos de dados) e quase 10%, *articles* (artigos). Para eles, a ausência de controle e qualidade dos dados era óbvia (ROUSIDIS *et al*., 2014, p. 285, tradução nossa).

Segundo os autores, uma abundância de problemas com mau uso dos dados foi identificada; problemas que constituem os dados inapropriados a propósitos de mineração. Eles indicam que mecanismos de prevenção a tais inconsistências devem ser criados (ex.:

uso de ID, como ORCID, na identificação de criador e contribuidor, formato de data padronizado, validação de data, entre outros). Em relação aos problemas provenientes do elemento *Type*, os autores acreditam que listas de valores pré-definidos (tipos de arquivos) podem ser úteis a autores durante preenchimento dos metadados. Concluem que o controle de dados faria repositórios mais atraentes e sustentáveis (ROUSIDIS *et al*., 2014, p. 285, tradução nossa).

2.2.6 Schopfel e outros (2014)

Para Schopfel e outros (2014), vincular dados a documentos é crucial para a interconexão de conhecimento científico. Eles colocam que enquanto editores acadêmicos fazem uso de novas tecnologias para enriquecer o conteúdo e as funcionalidades de seus produtos *online*, universidades deixam de aproveitar a oportunidade dos arquivos suplementares publicados com teses e dissertações eletrônicas (SCHOPFEL *et al*., 2014, p. 612-613, tradução nossa).

O artigo explora dados científicos relacionados a teses e dissertações eletrônicas como uma parte específica da emergente e-infraestrutura de pesquisa. Seu objetivo foi inserir teses de doutorado em um mundo de ciência digital aberta, e aprimorar acesso ao conhecimento científico e resultados de pesquisa financiada por setor público (SCHOPFEL *et al*., 2014, p. 613, tradução nossa).

De acordo com os autores, atualmente mais e mais resultados científicos são disseminados como conjuntos de dados em formatos digitais, às vezes conectados a ou em competição com publicações (SCHOPFEL *et al*., 2014, p. 613, tradução nossa).

Sistemas computacionais, dados, recursos informacionais, *networking*, sensores ativados digitalmente, instrumentos, organizações virtuais, observatórios, serviços e ferramentas interoperáveis por *software* – esses são os componentes tecnológicos de ciberinfraestrutura definidos pelo *US National Science Foundation Cyberinfrastructure Council* em 2007 (SCHOPFEL *et al*., 2014, p. 613, tradução nossa).

Com os dados, publicações se tornam documentos vivos, janelas aos resultados científicos. Documentos convencionais como artigos, teses, relatórios e anais de eventos são explorados como fonte primária de dados para mineração de texto, extração automática de informação, etc. Publicações de artigos podem servir como bases de dados para consultas cruzadas em literatura de domínio. Texto e arquivos de dados estão interconectados, porém possuem dois distintos conjuntos de metadados e são arquivados em servidores diferentes (SCHOPFEL *et al*., 2014, p. 614, tradução nossa).

Schopfel e outros (2014) afirmam que a publicação de conjuntos de dados em repositórios e sua disponibilização a revisores em pares atendem às estratégias da ciência

aberta e do acesso livre e irrestrito a resultados de pesquisa pública, pois garantem padrões de qualidade e permitem o reúso dos dados (SCHOPFEL *et al*., 2014, p. 615, tradução nossa).

No passado, teses e dissertações impressas eram submetidas com materiais suplementares em vários formatos e diferentes suportes (anexo impresso, cartão perfurado, disquete, fita de áudio, *slide*, CD-ROM, entre outros), o que dificultava seu processamento (localização no acervo) e reúso. Na nova infraestrutura de teses e dissertações eletrônicas, esses materiais podem ser submetidos e processados com os arquivos de texto. Se disseminados via repositórios abertos, esses resultados de pesquisa poderiam se tornar uma rica fonte de conjuntos de dados científicos, para reúso e outras explorações. Esses materiais complementares são geralmente *small data* ou *little science*, dados escondidos e inexplorados, de financiamento público e produção pessoal. Sua larga variedade afeta sua acessibilidade, abertura e reusabilidade (SCHOPFEL *et al*., 2014, p. 616, tradução nossa).

Os autores apontam que em muitos casos os dados são entregues como anexo das teses e dissertações depositadas em repositórios, o que impede seu reúso (dados em arquivos de texto formato PDF). Em alguns casos, os arquivos de texto e de dados são submetidos separadamente, tornando a tese ou dissertação porta de acesso aos dados (SCHOPFEL *et al*., 2014, p. 616, tradução nossa).

Schopfel e outros (2014) reiteram que fornecer acesso a dados científicos relacionados a teses e dissertações digitais é um desafio para bibliotecas acadêmicas, e com isso, fazem três questionamentos: “Qual sistema de informação melhor atende a tais necessidades? Como facilitar a recuperação desses conjuntos de dados? Quais são as condições legais para sua disseminação, acesso e reúso? (SCHOPFEL *et al*., 2014, p. 618, tradução nossa).

Repositórios de dados podem ser institucionais, como a maioria dos repositórios de teses e dissertações, porém também gerenciados por provedores terceirizados como o *Dryad*, *Zenodo* ou *Figshare*. Ademais, conjuntos de dados heterogêneos não podem ser comparados ao tipo de *Big Data* produzido pelo CERN e outros, pois são similares a dados pessoais. A arquitetura ideal deveria combinar características de armazéns de dados pessoais (*small data*) com aquelas de sistemas institucionais de informação (*big data*). Por conta da natureza específica dos dados e arquivos suplementares, parece apropriado não armazenar texto e arquivos de dados no mesmo repositório, mas distinguir entre servidores de documentos e repositórios de dados, depositando texto e dados em plataformas diferentes (SCHOPFEL *et al*., 2014, p. 618, tradução nossa).

Metadados apropriados são cruciais para a gestão de resultados de pesquisa e o desenvolvimento de repositórios e outros serviços. Sem eles, os dados permanecem escondidos. Pelas suas especificidades, conjuntos de dados, teses e dissertações deveriam

ser descritos por conjuntos de metadados diferentes (SCHOPFEL *et al*., 2014, p. 619, tradução nossa).

Padrões são necessários a interconexões e interoperabilidade. No caso das teses e dissertações e dados relacionados, isso significa metadados padronizados para ambos. Esses metadados devem ser documentados com os dados científicos e armazenados com a descrição dos requisitos técnicos. Segundo os autores, a falta de nomes permanentes para dados arquivados e da atribuição de identificadores persistentes únicos de objetos (URN, *Handle*, DOI) a conjuntos de dados científicos ainda é um problema (SCHOPFEL *et al*., 2014, p. 620, tradução nossa).

Os conjuntos de dados deveriam estar conectados aos textos via metadados das teses e dissertações, não somente pela atribuição de identificadores únicos a tais documentos ou por endereço URL, mas pelos seus próprios identificadores persistentes. Esses trabalhos acadêmicos deveriam mencionar a existência de conjuntos de dados relacionados como um elemento específico de seus conjuntos de metadados. Formatos, padrões, identificadores e apropriada vinculação são condições necessárias à recuperação de dados, contudo não garantem acessibilidade e reusabilidade. O manuseio e a integração de dados demanda serviço específico e gestão de fluxo de trabalho (SCHOPFEL *et al*., 2014, p. 620, tradução nossa).

Os dados depositados na forma de arquivos suplementares devem ser reusáveis independentemente da tese relacionada. Legalmente, eles também são independentes. Licenciamento não adaptado ou proteção em excesso por direitos autorais podem ser barreiras legais ao depósito, disseminação e reúso dos dados. A vinculação de conjuntos de dados à proteção *copyright* de teses e dissertações cria um potencial conflito com as políticas de dados abertos (SCHOPFEL *et al*., 2014, p. 621, tradução nossa).

Dados pessoais e informação estratégica ou sensível devem ser protegidos. Não obstante, Schopfel e outros (2014) apontam conflitos entre as políticas de acesso aberto e o interesse pessoal de pesquisadores (medo da concorrência e do plágio). Para eles, autor e instituição devem reconsiderar condições legais ao depósito e disseminação de conjuntos de dados, aplicando políticas de dados abertos. Sob sua ótica, documentos e dados devem ser distinguidos e separados intelectual, lógica e fisicamente (SCHOPFEL *et al*., 2014, p. 621, tradução nossa).

Investimentos foram feitos para facilitar a transição de teses impressas a digitais. Por Schopfel e outros (2014), novos investimentos são necessários à curadoria de dados científicos produzidos e depositados com teses e dissertações (SCHOPFEL *et al*., 2014, p. 622, tradução nossa).

Os autores entendem que a curadoria, a recuperação e o reúso dos dados seriam facilitados se esses fossem separados dos arquivos de texto científico. Todo desenvolvimento

deve ser padronizado e baseado no protocolo OAI. Portanto, repositórios de *small data* devem ser integrados em ambientes CRIS[14] (SCHOPFEL *et al*., 2014, p. 623, tradução nossa).

Para melhorar o acesso ao conhecimento científico, deve-se: avaliar as necessidades específicas da comunidade de pesquisa local; criar consciência por meio de debates e comunicação sobre o interesse público e científico desses dados; aprender com base em modelos e iniciativas de sucesso, intercambiar com *experts* e *stakeholders*; preparar um plano de gestão de dados científicos apropriado e adaptado às necessidades da comunidade acadêmica; e facilitar o acesso aberto a teses e dissertações como condição fundamental a um plano de gestão de dados relacionado (SCHOPFEL *et al*., 2014, p. 623, tradução nossa).

2.2.7 Mayernik (2015)

Em seu relato de pesquisa, o autor apresenta uma análise a favor da possibilidade das ciências climáticas terem tanto fortes, quanto fracas culturas de compartilhamento de dados. Ele considera práticas de dados o trabalho envolvido em criar, gerenciar, e usar dados científicos e seus metadados associados. Olhar para os desafios da curadoria de uma perspectiva institucional pode providenciar uma estrutura transversal para analisar, entender, e facilitar tais práticas dentro e através das organizações, disciplinas, e projetos. Para Mayernik (2015), instituições são configurações sociais que fornecem estabilidade e significado ao comportamento humano. Elas reprimem e habilitam a forma como as pessoas pensam, agem, e cooperam (MAYERNIK, 2015, p. 973-974, tradução nossa).

O autor distingue os conceitos de “instituição” e “organização”, onde “instituição” está para algo abstrato. Práticas de dados científicos estão igualmente incorporadas em múltiplas instituições. Em seu artigo, ele apresenta uma análise teórica que traça 5 características institucionais centrais ao entendimento das práticas de dados científicos: (a) normas e símbolos, (b) intermediários, (c) rotinas, (d) padrões, e (e) objetos materiais. Elas são discutidas como “suportes institucionais” para a curadoria de dados. Mayernik (2015) aplica esse quadro teórico a três estudos de caso de práticas científicas – *Center for Embbeded Networked Sensing* (CENS), *Long Term Ecological Research* (LTER) e *University Corporation for Atmospheric Research* (UCAR) – a fim de ilustrar formas em que suporte institucional à gestão de dados varia dentro de uma organização ou disciplina acadêmica (MAYERNIK, 2015, p. 974, tradução nossa).

O objetivo do artigo foi avançar o entendimento de práticas de dados científicos de 3 maneiras: (a) desenvolver um quadro para descrever as características institucionais relevantes a práticas de dados científicos; (b) ilustrar como tomar uma disciplina como unidade

[14] *Current Research Information System* ou Sistema de Informação de Pesquisa Atual.

de análise, enquanto informativo, encobre o espectro de práticas que existem em cada uma delas; e (c) ilustrar como fatores críticos ao redor de práticas de dados podem ser caracterizados por um número de interrelações entre suportes institucionais (MAYERNIK, 2015, p. 974, tradução nossa).

Para o autor, relatórios de alto nível sobre coleções de dados científicos geralmente enfatizam diferenças disciplinares, porém estão menos propensos a discutir como ou se existe variabilidade similar em tais disciplinas (MAYERNIK, 2015, p. 974, tradução nossa).

Sob sua ótica, instituições são padrões estáveis de comportamento humano que estruturam, legitimam, ou deslegitimam ações, relacionamentos, e entendimentos dentro de situações sociais particulares. Organizações que desenvolvem sistemas e serviços de curadoria de dados estão no nexo de múltiplos tipos de instituições. Acadêmicos têm utilizado análises baseadas em instituições para estudar problemas de dados e de informação (MAYERNIK, 2015, p. 975, tradução nossa).

Instituições de pesquisa científica são repletas de expectativas normativas sobre como o trabalho científico deve ser conduzido, compartilhado, e avaliado. Uma ilustração do componente simbólico de instituições de pesquisa acadêmica é a atividade comum da criação de uma logo que represente um projeto ou organização. No entanto, símbolos não precisam ser visuais. Políticas diretivas podem ter fortes papéis simbólicos, como a exigência de plano de gestão de dados em propostas de pesquisa, pois tal medida pode não ter efeito prático (MAYERNIK, 2015, p. 976-977, tradução nossa).

Intermediários podem ser indivíduos ou organizações. Eles têm papéis críticos em possibilitar dados científicos a serem compartilhados, entendidos, e usados. O autor aponta para mudanças institucionais que tais intermediários vêm sofrendo com o avanço da tecnologia. Para ele, as rotinas constituem uma parte importante da vida cotidiana, assim como o uso/criação de dados e informação. Entretanto, elas não são definidas por um conjunto formalizado de passos. Quanto aos padrões, ele afirma que são formulários institucionais altamente codificados que carregam regras e especificações de como organizar, formatar, documentar, e consolidar informação e outras entidades. O desenvolvimento de padrões é uma atividade comum em ciências da informação e de dados. Ele coloca que negociações entre *stakeholders* são necessárias para que se obtenha alinhamento entre padrões e práticas de trabalho (MAYERNIK, 2015, p. 977-978, tradução nossa).

Mayernik (2015) afirma que as instituições são carregadas por objetos materiais assim como por humanos. Documentos digitais, *software*, ou dados forçam as formas como ferramentas, padrões, e abstrações computacionais têm se desenvolvido e perdurado por décadas. Apesar de revoluções tecnológicas ocorrerem frequentemente, mudanças institucionais relacionadas a tais tecnologias são tipicamente evolucionárias por natureza.

Objetos materiais podem carregar suposições e expectativas sobre padrões de comportamento de uma situação a outra, como de desenvolvedores de sistemas de informação a usuários (MAYERNIK, 2015, p. 978, tradução nossa).

Para o autor, seu quadro de cinco suportes institucionais esboça comportamentos e processos sociais propícios ao estabelecimento da curadoria de dados como uma atividade institucional (MAYERNIK, 2015, p. 978, tradução nossa).

Ao final das avaliações feitas nos estudos de caso, o autor compreende que:

1. As infraestruturas de dados do CENS eram muito diferentes em cada projeto, e que a gestão de dados e a preservação em longo prazo não eram apoiados dentro dos projetos focados na ciência de campo, incluindo projetos de ecologia. Padrões de dados e metadados de domínio ou não eram úteis, ou mal interpretados;

2. Os gestores de informação da LTER providenciam um quadro de intermediários treinados para dar suporte e facilitar as rotinas que são necessárias à existência e comparabilidade dos dados ao longo do tempo. A rede LTER possui um padrão oficial de metadados chamado *Ecological Metadata Language* [15] (EML), que embora de difícil implementação, está fortemente institucionalizado;

3. O suporte técnico e organizacional para a gestão, o compartilhamento, e a preservação de dados científicos varia consideravelmente na UCAR. Políticas e planos de alto nível da UCAR promovem uma normativa e simbólica cultura de compartilhamento de dados. Existem grupos de gestão de dados que fornecem serviços de alto nível técnico na organização, contudo muitos projetos não recebem o mesmo apoio profissional por conta de seu tamanho, limitações de financiamento ou escopo (MAYERNIK, 2015, p. 981-986, tradução nossa).

Mayernik (2015) conclui que as cinco categorias de suporte institucional esboçadas em seu artigo provêm um meio para analisar como as práticas de gestão, curadoria, e preservação de dados emergem e evoluem dentro e através de instituições científicas. Aplicá-las a diferentes cenários de pesquisa ilustra como práticas de dados por parte dos pesquisadores variam entre disciplinas. Ele afirma que a computação e os dados digitais providenciam novos meios de condução da pesquisa científica, e que lidar com tais dados é uma tarefa complexa. Para o autor, seu artigo sugere que para desenvolver e manter a curadoria de dados e metadados como atividade autocontínua, objetivos e relacionamentos institucionais precisam mudar de forma a reconhecer as múltiplas facetas da curadoria de dados como uma questão institucional (MAYERNIK, 2015, p. 989, tradução nossa).

[15] Uma especificação de metadados desenvolvida pela e para a disciplina de Ecologia, baseada em trabalho prévio feito pela Sociedade de Ecologia da América e esforços associados. A EML é implementada como um esquema XML que pode ser usado para documentar dados ecológicos.

Por fim, indica que o sucesso de novas estruturas institucionais à curadoria de dados dependerá da correlação positiva entre resultados específicos desejados, como por exemplo um melhor e mais fácil arquivamento e uso de conjuntos de dados científicos, resultados encontrados por meio desse uso, etc. (MAYERNIK, 2015, p. 990, tradução nossa).

2.2.8 Amorim e outros (2016)

O trabalho apresenta uma visão geral de várias plataformas de gestão de dados científicos que podem ser implementadas por uma instituição. Primeiro, identifica repositórios conhecidos que estão sendo utilizados para a gestão de publicações e de dados simultaneamente. Para o levantamento, utilizou-se a base indexadora de repositórios intitulada *OpenDOAR*. A pesquisa buscou conhecer a adequação desses repositórios no manuseio de dados científicos (requisitos de metadados de domínio e orientações à preservação). (AMORIM *et al*., 2016, p. 851, tradução nossa).

Para Amorim e outros (2016), enquanto publicações podem ser precisamente descritas por bibliotecários, metadados de qualidade à descrição de conjuntos de dados requerem a contribuição dos pesquisadores envolvidos em sua produção, isso devido a seu conhecimento de domínio. Eles apontam para plataformas estagiárias, as quais são adaptadas para capturar registros de metadados enquanto esses são produzidos. Essas plataformas são capazes de exportar conjuntos de dados e registros de metadados a repositórios de dados científicos. Juntos, os repositórios de dados e tais plataformas providenciam ferramentas para lidar com os estágios do fluxo de trabalho científico (AMORIM *et al*., 2016, p. 852, tradução nossa).

A prática de citação de dados ainda é incomum, mas está crescendo. Amorim e outros (2016) mencionam os artigos de dados, publicações que servem como contexto e referência aos dados. Para os autores, muitos dos desafios de *design* e desenvolvimento de repositórios institucionais estão na descrição e na preservação dos dados científicos em longo prazo. Além de capturar metadados durante processo de importação, os repositórios de dados distribuem informação a outras instâncias, melhorando a visibilidade das publicações por meio de motores de busca especializados e bases indexadoras de repositórios. Instituições governamentais também estão promovendo a divulgação de dados abertos para aprimorar a transparência pública, o que motiva o uso de plataformas de gestão de dados (AMORIM *et al*., 2016, p. 852, tradução nossa).

Os autores observam que repositórios como o *DSpace* são amplamente utilizados entre instituições com vistas à gestão de publicações, e que essas instituições podem apoiar a plataforma a se expandir, atendendo a requisitos adicionais. Pontuam que alguns repositórios não implementam interfaces com indexadores de repositórios, o que poderia

influenciar a atualização estatística nas bases indexadoras (AMORIM *et al*., 2016, p. 853, tradução nossa).

Serviços prestados pelo EUDAT, *Figshare* e *Zenodo* consistem em uma instalação única que recebe todos os dados depositados, ao invés de uma matriz distribuída de instalações gerenciáveis. Plataformas apoiadas por governo (ex.: CKAN) estão sendo utilizadas como parte de iniciativas de abertura governamental em diversos países, permitindo transparência a dados sensíveis como execução de orçamento, etc. Para os autores, o acesso ao código-fonte pode ser um critério valioso na seleção de uma plataforma, evitando assim problemas de descontinuidade de determinado serviço. A disponibilidade do código-fonte permite também modificações adicionais (fluxos de trabalho personalizados). Ademais, eles entendem que a existência de uma *Application Programming Interface*[16] (API) possibilita manutenção e futuro desenvolvimento do repositório. Percebem que algumas plataformas falham ao não fornecerem identificadores únicos a recursos depositados, o que dificulta a citação dos dados em publicações (AMORIM *et al*., 2016, p. 853, tradução nossa).

Os autores compilam em uma tabela os repositórios investigados e suas limitações. Nela, destacam-se o CKAN e o *Omeka*, pois ambos não atribuem identificadores únicos e não estão em conformidade com o OAI-PMH.

Enquanto principais provedores de metadados, pesquisadores são responsáveis pela descrição de seus dados científicos. Logo, têm papel fundamental no processo de depósito dos dados. Além disso, os autores acreditam que instituições valorizam metadados em conformidade com padrões, o que faz com que seus dados estejam prontos para inclusão em ambientes em rede, assim aumentando sua visibilidade. Para Amorim e outros (2016), os curadores de dados são geralmente especialistas em informação, e que se espera que uma estreita colaboração com pesquisadores possa resultar em registros de metadados detalhados e condescendentes. Os coletores de dados podem ser tanto indivíduos, quanto serviços que indexam o conteúdo de repositórios (AMORIM *et al*., 2016, p. 854, tradução nossa).

Indicam o OAIS, o PREMIS e o METS como padrões para a preservação em longo prazo, porém pontuam que tais padrões são de difícil instalação e manutenção quando adotados por instituições na longa cauda da ciência – instituições de baixo ou médio orçamento que criam grandes números de pequenos conjuntos de dados (AMORIM *et al*., 2016, p. 855, tradução nossa).

Para eles, uma instituição pode tanto terceirizar um serviço externo quanto instalar e personalizar seu próprio repositório (assistindo custos de manutenção). Afirmam que o *DSpace*, o *ePrints*, o CKAN ou qualquer solução Fedora podem ser instalados e executados

[16] Interface de Programação de Aplicações.

sob controle da instituição de pesquisa (melhor controle sobre dados arquivados). (AMORIM et al., 2016, p. 855, tradução nossa).

O *ePrints* e o *DSpace* não são projetados para assistir ambientes colaborativos em tempo real, onde pesquisadores podem produzir e descrever seus dados incrementalmente. Adotar abordagens dinâmicas à gestão de dados pode motivar pesquisadores a usarem plataformas de gestão como parte de sua atividade de pesquisa diária, enquanto trabalham nos dados (AMORIM *et al*., 2016, p. 856, tradução nossa).

De acordo com Amorim e outros (2016), o *DSpace*, o *ePrints*, o *Zenodo*, e o EUDAT permitem períodos específicos de embargo. Depois de expirados, os dados são disponibilizados à comunidade. O CKAN e o *Figshare* apenas possuem opção de arquivamento privado dos dados. Além disso, o CKAN é o único incompatível com o OAI-PMH, portanto, não possibilita exportação em esquemas de metadados. Apontam o CKAN como uma iniciativa originalmente projetada a dados governamentais. Nenhuma das plataformas investigadas suporta estágios de validação colaborativa, onde curadores e pesquisadores impõem a correta estrutura de dados e metadados. Para os autores, é interessante não apenas avaliar plataformas de acordo com a facilidade de descoberta por máquinas, mas também observar com que facilidade os humanos podem encontrar conjuntos de dados nelas (AMORIM *et al*., 2016, p. 857, tradução nossa).

O CKAN possui muitos casos de sucesso com dados governamentais, embora em falta de cenários relacionados à gestão de dados científicos. O *DSpace*, conhecido por sua capacidade de lidar com publicações de pesquisa, tem sido reconhecido também por manusear dados científicos. O *Zenodo* é uma solução à longa cauda da ciência apoiada pelos laboratórios do CERN (AMORIM *et al*., 2016, p. 858, tradução nossa).

Há falta de apoio à captura de dados em estágios iniciais de atividades de pesquisa. A introdução antecipada da descrição e do depósito dos dados no fluxo de trabalho da pesquisa significa que as descrições estarão parcialmente prontas ao final da coleta de dados. Pesquisadores não são especialistas em gestão de dados, então necessitam de ferramentas efetivas que os permitam produzir registros de metadados padronizados sem que tenham que aprender tais padrões (AMORIM *et al*., 2016, p. 858, tradução nossa).

Para os autores, a avaliação mostrou que pode ser difícil selecionar uma plataforma sem primeiro estudar os requisitos de todos *stakeholders*. Destacam que, embora o CKAN seja primariamente utilizado para dar transparência a dados governamentais, seus recursos e API o tornam apto à gestão de dados científicos. Algumas instituições talvez queiram os servidores onde os dados são armazenados sob seu controle, assim como gerenciar diretamente seus conjuntos de dados. Plataformas como o *DSpace* ou o CKAN são apropriadas para tal, pois podem ser instaladas em um servidor institucional (AMORIM *et al*., 2016, p. 860-861, tradução nossa).

2.2.9 Beaujardière (2016)

O artigo discute atividades da *United States' National Oceanic and Atmospheric Administration* (NOAA) que dão suporte à gestão de dados ambientais. Inicialmente, apresenta uma visão conceitual do estado das atividades de gestão de dados que se deseja alcançar. Também descreve um quadro de políticas de gestão de dados ambientais, práticas organizacionais, e considerações técnicas para assistir o acesso contínuo e efetivo a observações da Terra e produtos derivados (BEAUJARDIÈRE, 2016, p. 6, tradução nossa).

O NOAA *Environmental Data Management Framework* [17] busca promover entendimento de políticas e atividades de gestão de dados; maximizar a probabilidade de dados ambientais serem encontráveis, acessíveis, bem documentados, e preservados para uso futuro; e encorajar o desenvolvimento e o uso de ferramentas e práticas uniformes para o manuseio de dados ambientais internamente. Os elementos básicos do quadro são: Princípios, Governança, Recursos, Padrões, Arquitetura, e Avaliação. Para o autor, tais elementos se aplicam a muitas classes de dados, e a ciclos de vida de dados que podem ser específicos a coleções de dados particulares (BEAUJARDIÈRE, 2016, p. 8-9, tradução nossa).

O ciclo de vida dos dados definido pelo autor inclui todas as atividades que afetam um conjunto de dados antes e durante seu tempo de vida. Distintos conjuntos de dados podem ter diferentes tempos de vida. O uso do termo "ciclo de vida" inclui preservação em longo prazo e não implica uma vida útil finita ou um período limitado de utilidade. Seu modelo de atividades no ciclo de vida dos dados se divide em três grupos: Planejamento e Produção, todas as atividades até e incluindo o momento em que uma observação é capturada; Gestão dos Dados, atividades que envolvem processar, verificar, documentar, publicitar, distribuir, e preservar os dados; e Uso, todas as atividades realizadas pelo consumidor dos dados, como descobrir, analisar, e citar os dados (BEAUJARDIÈRE, 2016, p. 9-10, tradução nossa).

O ciclo de vida dos dados é um processo dinâmico, cujos passos são interdependentes (ex.: documentação inadequada em fase inicial pode impedir uso posterior). Os dados podem passar por múltiplos ciclos de uso e reúso, onde dados originais geram dados derivados, que também precisam ser coletados e gerenciados (BEAUJARDIÈRE, 2016, p. 10, tradução nossa).

Quanto aos Princípios, eles são: Acesso Total e Aberto; Preservação em Longo Prazo; Qualidade da Informação; e Fácil Uso. Em geral, dados gerenciados ou pagos por financiamento público deveriam estar disponíveis ao público assim que possível após sua

[17] Quadro de Gestão de Dados Ambientais do NOAA.

coleta, de forma indiscriminada, a custo mínimo (BEAUJARDIÈRE, 2016, p. 11, tradução nossa).

Os dados deveriam ser ofertados em formatos conhecidos por trabalhar com uma ampla variedade de ferramentas científicas ou de apoio à decisão. Vocabulários comuns, semântica, e modelos de dados deveriam ser empregados (BEAUJARDIÈRE, 2016, p. 13, tradução nossa).

Dados não podem ser adequadamente gerenciados sem recursos, incluindo pessoal, orçamento, e outros elementos de apoio. A falta desses recursos é geralmente um fator que leva a dados mal documentados, inacessíveis, e/ou indevidamente preservados (BEAUJARDIÈRE, 2016, p. 15, tradução nossa).

Aprimoramentos em gestão de dados não podem ser feitos com base em esforços voluntários. Atividades tais como criar e conservar metadados, tornar dados disponíveis a usuários, ou assegurar que dados sejam transmitidos a uma acomodação arquivística deveriam ser incluídas entre obrigações regulares de equipes (BEAUJARDIÈRE, 2016, p. 16, tradução nossa).

Diferentes tipos de padrões são aplicáveis em várias fases do ciclo de vida dos dados. Esses incluem vocabulários, padrões de qualidade dos dados, padrões de metadados que especificam o conteúdo e a estrutura da documentação sobre um conjunto de dados, modelos de dados e padrões de formato que detalham o conteúdo e a estrutura dos próprios dados digitais, e padrões de interface que especificam como os serviços são invocados. Alguns padrões são de propósito geral e podem requisitar especialização para tipos particulares de dados. A adoção de padrões comuns dá suporte à interoperabilidade, a qual permite diversos dados, ferramentas, sistemas, e arquivos serem combinados sem que seja necessário escrever *software* personalizado para lidar com todos os *links* entre dados (BEAUJARDIÈRE, 2016, p. 16-17, tradução nossa).

Beaujardière (2016) afirma que é mais eficiente tornar um conjunto de dados acessível a partir de uma única fonte autorizada a fazer um usuário baixar, manter, e possivelmente redistribuir múltiplas cópias (BEAUJARDIÈRE, 2016, p. 18, tradução nossa).

Aprimoramentos de metadados incluem a uniformização de vocabulários e atalhos XML a informações-chave, assim como a desambiguação entre conjuntos de dados que possuem títulos idênticos, porém que diferem em tempo ou em outros atributos (BEAUJARDIÈRE, 2016, p. 21, tradução nossa).

O DOI pode ser utilizado como uma referência à localização atual dos dados. Ele também é persistente, o que significa que uma vez atribuído, jamais pode ser deletado ou reatribuído (BEAUJARDIÈRE, 2016, p. 21, tradução nossa).

Nas conclusões, o autor sugere a qualquer organização com o intuito de gerenciar dados ambientais que: escreva planos de gestão de dados e distribua uma porcentagem de

fundos ao gerenciamento de dados resultantes; solicite *feedback* dos usuários quanto à acessibilidade, usabilidade, e qualidade dos dados; use dados, metadados, e padrões de protocolo domésticos e internacionais sempre que adequado; estabeleça uma capacidade de busca federada através de múltiplos catálogos e fontes de metadados que podem ser consultados por usuários de dados e por catálogos externos ou temáticos; e assegure que a equipe responsável por dados ambientais entenda a necessidade de sua gestão e que seja treinada em boas práticas de gestão de dados ambientais (BEAUJARDIÈRE, 2016, p. 25, tradução nossa).

2.2.10 Gómez, Méndez e Hernández-Pérez (2016)

Gómez, Méndez e Hernández-Pérez (2016) afirmam que a abertura de dados científicos é recomendada pela OECD, requisitada pelo governo estadunidense e várias agências de fomento como a *National Science Foundation* (NSF) e a *National Institutes of Health* (NIH). Para os autores, o compartilhamento dos dados: aumenta a possibilidade da pesquisa obter maior impacto e visibilidade; favorece a reprodutibilidade da ciência; economiza gastos ao gerar dados; promove colaboração; e contribui com o aumento da credibilidade no sistema (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 546-547, tradução nossa).

Para cada domínio científico há uma interpretação particular sobre conjuntos de dados científicos, sua natureza, procedimentos de coleta, e descrição de metadados. Nas Ciências Sociais e Humanas (CSH) nem todos os dados são coletados digitalmente, assim como dados podem assumir outras formas e formatos. Outro problema nas CSH está na fonte dos dados, pois muitas investigações são baseadas em dados que não foram originalmente produzidos por ou para pesquisadores. Os pesquisadores das CSH geram menos dados por observações, pois geralmente tendem a usar dados de todos os tipos de fontes (sons, filmes, jornais, registros administrativos, entre outros). Dentro do espectro de assuntos e disciplinas que cobrem as humanas pode haver diferentes definições sobre dados, o que complicaria a perspectiva de seu gerenciamento e recuperação (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 547, tradução nossa).

Dados científicos apresentam uma heterogeneidade que varia radicalmente entre disciplinas, áreas temáticas, grupos de pesquisa, e pesquisadores. Agências de fomento estão sensibilizando e colocando pressão para que pesquisadores gerenciem, compartilhem, facilitem a recuperação e a preservação de seus dados, além de assegurar que sejam FAIR (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 547, tradução nossa).

De acordo com os autores, quando pesquisadores compartilham seus dados e metadados em um repositório de dados, eles deveriam traduzir a metainformação que utilizam

em seus ambientes virtuais de pesquisa em esquemas de metadados usados no repositório. Em seu estudo, eles analisam os esquemas de metadados usados por repositórios. Abordam o problema da gestão de dados científicos nas CSH por meio de um estudo de repositórios de dados nesses domínios. Os repositórios investigados estão indexados na base RE3DATA, indicada como registro referência para repositórios de dados, recomendada pela *European Comission*, *PeerJ*, *Springer*, *Nature's Scientific Data*, etc. (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 547-548, tradução nossa).

Das questões de pesquisa elaboradas pelos autores, destacam-se: quais tipos de dados são armazenados e gerenciados por repositórios das CSH? Quais esquemas de metadados são utilizados nesses repositórios? Há um esquema predominante em cada caso? Seus objetivos foram: identificar os repositórios de dados científicos nas CSH; estudar quais tipos de dados resultam de pesquisas nesses domínios; e apresentar os esquemas de metadados mais utilizados nesses repositórios (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 548, tradução nossa).

Um dos critérios de seleção para a escolha dos repositórios estudados foi baseado na cobertura ou no número de conjuntos de dados armazenados neles.

Seus resultados indicaram que dados estatísticos e científicos (formatos como .spss, .fits, .gis, etc.), documentos (*Word*, *Excel*, ou formatos *openoffice* similares), e imagens (.jpeg, .gif, .tif, .png, .svg, etc.) são os mais comumente utilizados em projetos de digitalização nas Ciências Humanas (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 550, tradução nossa).

Os esquemas de metadados mais usados nos campos das CSH são o *Dublin Core* e a DDI (*Data Documentation Initiative*). Um quarto dos repositórios das CSH indexados na RE3DATA declararam algum tipo de esquema de metadados. Desses, 45% usam *Dublin Core*. Nota-se pequena incidência do uso da EML nas CSH. Isso ocorre porque repositórios multidisciplinares podem utilizar mais de um esquema de metadados simultaneamente (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 551, tradução nossa).

Entre os 6 repositórios selecionados à análise detalhada, os 3 das Ciências Sociais têm o DC e a DDI como esquemas dominantes. Os outros 3 das humanas apresentam significante variância entre esquemas, contudo com presença unânime do DC. Observa-se que os repositórios *UK Data Service* (Ciências Sociais) e *Prometheus* (Humanas) também utilizam o METS (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 551-552, tradução nossa).

A heterogeneidade e a complexidade dos repositórios de dados científicos são manifestadas nos esquemas de metadados escolhidos para descrever os dados, ainda mais evidente nas humanas. Em seus resultados, os autores afirmam que o RE3DATA atingiu excelência na busca por repositórios de dados científicos em todas as disciplinas. Um dos

problemas encontrados era a impossibilidade de atualizar informações dos repositórios indexados de forma *online*, sendo preciso preencher e encaminhar formulário ao RE3DATA (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 552, tradução nossa).

Concluem que o RE3DATA confirma a tendência do uso da DDI nas Ciências Sociais, e que isso pode ser motivado pela maturidade do padrão (criado em 1995), ou por ele ter sido originalmente desenvolvido para descrever dados, não documentos. Relatam que nas Ciências Humanas foi observado o uso de esquemas próprios ou adaptados a partir do DC, e que não é incomum ver outros padrões de metadados (EDM, METS, e MIDAS) sendo utilizados na descrição de dados das humanas em seus repositórios. Salientam que o DC é um padrão *default* destinado a repositórios de publicações, e que tal tendência abrange repositórios de dados, ao menos em primeira instância. Todavia, afirmam que ainda é cedo para confirmar se ele pode ser adaptado às peculiaridades de todos os dados científicos disciplinares (GÓMEZ; MÉNDEZ; HERNÁNDEZ-PÉREZ, 2016, p. 553-554, tradução nossa).

2.2.11 Garnett e outros (2017)

De acordo com Garnett e outros (2017), o acesso a dados científicos no Canadá depende de padrões, repositórios, e redes que dão pouca atenção à descoberta interdisciplinar. Oportunidades são limitadas por padrões de metadados de domínio. A descoberta de dados é definida no artigo como a habilidade de obter novas informações e conhecimento a partir de fontes de dados existentes. A descoberta depende de metadados consistentes que estruturam informação sobre dados científicos. Os autores apontam que à luz dos Princípios FAIR, dados científicos deveriam ser estruturados de modo a facilitar sua descoberta por humanos e máquinas. Sem dados FAIR, a descoberta e o reúso se tornam difíceis, pois um único pesquisador pode ter que ir a vários lugares para encontrar e acessar dados (GARNETT *et al*., 2017, p. 201-202, tradução nossa).

O artigo reporta o desenvolvimento de um serviço de descoberta nacional canadense para dados científicos por meio da rede *Portage*, estabelecida em 2015 pela *Canadian Association of Research Libraries* (CARL). A *Portage* tem o objetivo de desenvolver uma rede de *expertise*, infraestrutura, e ferramentas de apoio a pesquisadores na preparação, curadoria, preservação, e descoberta de seus dados científicos. Os autores discutem oportunidades e desafios que emergem dos padrões de metadados de domínio e o desenvolvimento de um serviço de descoberta para dados científicos no Canadá (GARNETT *et al*., 2017, p. 202, tradução nossa).

Garnett e outros (2017) citam o projeto piloto inglês JISC *Research Data Discovery Service*, que colhe metadados utilizando o protocolo OAI-PMH, e que adota o *software* de repositório digital, CKAN; o serviço de descoberta de dados científicos da *Australian National*

Data Service (ANDS) chamado *Research Data Australia*, que colhe metadados de dados científicos em mais de cem instituições de pesquisa na Austrália; a base de dados acadêmica *SHARE*, cuja infraestrutura inclui um processo de colheita que usa o OAI-PMH e APIs abertas (foco no arquivamento e no enriquecimento de metadados); e o *Federated Research Data Repository* (FRDR), uma iniciativa entre a CARL/*Portage* e a *Compute Canada* que busca criar uma plataforma capaz de armazenar e servir grandes volumes de dados, assim como possibilitar a descoberta de dados científicos canadenses a partir de uma única interface de busca (GARNETT *et al*., 2017, p. 203-204, tradução nossa).

Os metadados que levam à descoberta de dados científicos são informações estruturadas sobre os próprios dados, autores, produtores, datas de coleta, entre outros. Os autores definem o *Dublin Core* como um padrão de metadados descritivos usado por repositórios digitais como o *DSpace*, que descreve objetos digitais incluindo dados científicos. Já o *DataCite* assiste a descoberta de dados científicos na *Web* ao focar em elementos que definem a localização, identificação, e citação única desses dados. O *DataCite* requisita a criação de DOIs, que permitem fácil identificação e citação de dados, e fornecem metadados persistentes, estejam os dados abertos ou não (GARNETT *et al*., 2017, p. 205-206, tradução nossa).

Em um quadro, os autores apresentam 11 padrões de metadados comumente utilizados em repositórios de dados científicos canadenses. Entre eles estão o *Darwin Core* (Biodiversidade), o DDI (genérico/Ciências Sociais), o *DataCite* (genérico), o EML (Ecologia), e a ISO 19115 (geográfico). A complexidade e a granularidade dos metadados assistidos por padrões de domínio variam muito. Alguns padrões de metadados de domínio são facilmente mapeados em padrões genéricos como o DC e o *DataCite*, embora tais padrões possuam especificidades que não deveriam ser sacrificadas. Por exemplo, metadados quanto a sistemas de coordenadas espaciais e escala são importantes para a descoberta e reúso de dados geoespaciais. Dessa maneira, a variação disciplinar se torna relevante à descoberta de dados científicos. Para os autores, embora seja necessário estar atento a padrões de domínio, nenhum padrão é perfeito. Assim, a flexibilidade desses padrões será requisito a futuros modelos de metadados (GARNETT *et al*., 2017, p. 206-207, tradução nossa).

Segundo Garnett e outros (2017), o portal de dados abertos do governo canadense utiliza uma implementação da plataforma CKAN, similar a outros governos. Por padrão, a maioria dos portais OAI atende a metadados *Dublin Core*. Isso pode ser adequado a buscas simples, mas não a pesquisas interdisciplinares avançadas. Metadados de domínio podem ser apresentados como diferentes facetas na interface de descoberta, permitindo a busca e a descoberta de metadados granulares (GARNETT *et al*., 2017, p. 208, tradução nossa).

Os autores citam o processo de mapeamento dos padrões de metadados encontrados em uso nos repositórios de dados canadenses e a adaptação de seus elementos aos do *Dublin Core*, de forma a alcançar parte do objetivo estabelecido ao projeto FRDR. Assim, relatam que enquanto alguns dos padrões de metadados de domínio selecionados foram fáceis de alinhar ao DC, outros requisitaram a procura de amostras de metadados adicionais em *Websites* de organizações ou em repositórios, antes de confirmar a adequação das opções específicas de mapeamento. O resultado desse processo enfatiza os elementos centrais entre todos os padrões e aponta para quão variados os metadados de domínio podem ser (GARNETT *et al*., 2017, p. 209-210, tradução nossa).

Dados científicos não existem independentemente, pois são relacionados a outras produções científicas como publicações, anais, e planos de gestão de dados. Garantir que dados científicos sejam descobertos requer considerar como são vinculados a fontes relacionadas. *Links* a publicações que citam um conjunto de dados devem ser criados precisamente, e com o objetivo de facilitar a descoberta entre plataformas, devem ser feitos por meio do uso de identificadores persistentes como o DOI. A maioria dos padrões de domínio e genéricos permitem referências a recursos associados, embora o uso de elementos referenciais seja raramente obrigatório a depositantes (GARNETT *et al*., 2017, p. 211-212, tradução nossa).

Metadados de dados científicos podem também ser descritos e armazenados como dados vinculados, expressos no formato *Resource Description Framework* (RDF), o qual usa vocabulários e ontologias para definir dados e suas ligações (GARNETT *et al*., 2017, p. 212, tradução nossa).

Ademais, dados científicos podem ser vinculados a instrumentos, padrões disciplinares e ontologias, repositórios de código e *software*, entre outros, o que aperfeiçoa significativamente a reprodutibilidade e reúso dos dados. Aprimorar o acesso a métricas sobre o compartilhamento e o reúso de dados científicos permite uma avaliação mais efetiva dos investimentos em pesquisa. Com o ORCID e o DOI, a harmonização entre sistemas de informação se torna mais provável por meio de automação, apesar de boa parte do trabalho de enriquecimento de metadados e de ligações seja manual (GARNETT *et al*., 2017, p. 213, tradução nossa).

O artigo resume o trabalho desempenhado no desenvolvimento de um modelo de metadados que se destina à descoberta de dados científicos. Conclui que o uso de padrões e metadados abertos por parte do FRDR tem melhorado a descoberta de dados científicos no Canadá (GARNETT *et al*., 2017, p. 215, tradução nossa).

2.2.12 Radio e outros (2017)

A proliferação de conjuntos de dados e sua disponibilidade em vários repositórios exigem metadados que forneçam contexto e clareza organizacional para possibilitar seu reúso. Entretanto, conjuntos de dados, como um tipo de coleção, surgem em inúmeras formas, estruturas, e relacionamentos. Como as características dos conjuntos de dados variam entre disciplinas, é razoável sugerir que os métodos pelos quais eles são descobertos sejam informados para diferenciar áreas de pesquisa. Para que sua unidade interna seja representada precisamente, é necessário que sua documentação por meio de metadados seja um reflexo ao longo dos mesmos espectros ônticos. Ao explorar diferentes padrões de metadados e acompanhar práticas em diferentes domínios, é possível identificar as estruturas que os sustentam, o que pode servir para facilitar maior clareza na organização e na descrição (RADIO *et al*., 2017, p. 161-162, tradução nossa).

Muitos padrões de metadados utilizados para descrever conjuntos de dados são construídos sobre algumas estruturas de dados. Como exemplo, a estrutura de árvore implica suposições sobre a organização do objeto que ela descreve, ou seja, sugere características hierárquicas aos conjuntos de dados, embora nem todos sejam organizados hierarquicamente (RADIO *et al*., 2017, p. 162, tradução nossa).

Para os autores, um dos desafios na descrição de conjuntos de dados científicos está na dificuldade em descrever o *software* associado a esses conjuntos. Documentá-lo se torna complicado pelo fato de que tal *software* está ontologicamente separado dos dados (RADIO *et al*., 2017, p. 162, tradução nossa).

O artigo examina estruturas de dados para metadados e suas implicações. Subsequentemente, práticas das Ciências da Vida, Ciências Sociais, Ciências Humanas, e Sistemas de Informação Geográfica são examinadas pelas suas habilidades em serem precisamente documentadas por padrões comuns (RADIO *et al*., 2017, p. 162, tradução nossa).

No contexto de repositórios onde metadados são comumente compartilhados, é desejável que os formatos em uso sejam amplamente interoperáveis, o que leva à dicotomia entre a multiplicidade de padrões e o baixo número de estruturas de dados que os enquadram. Conjuntos de dados em suas inúmeras estruturas organizacionais superaram rapidamente muitas das capacidades descritivas de seus padrões de metadados correspondentes (RADIO *et al*., 2017, p. 163, tradução nossa).

Os autores afirmam que seria complicado descrever os padrões DC como semanticamente ricos por conta da sua estrutura plana e sua dificuldade em contextualizar informação, especialmente pela perspectiva da máquina. Uma estrutura plana consiste em

um conjunto desordenado de relacionamentos *key:value* (RADIO *et al*., 2017, p. 164-165, tradução nossa).

Talvez a instância mais reconhecível de um modelo de árvore no contexto dos metadados seja o padrão XML, que enquanto não prescreve elementos a serem usados, providencia a estrutura na qual eles podem ser criados. O *Metadata Object Description Schema* (MODS), o *Ecological Metadata Language* (EML), e o *DataCite* são exemplos de padrões baseados em XML. Uma árvore é tecnicamente um gráfico não direcionado. Gráficos consistem em vértices e arestas, e podem ser direcionados ou não. Um exemplo de um gráfico direcionado seria o *Resource Description Framework* (RDF), uma linguagem projetada para descrever recursos por meio de expressões atômicas conhecidas como triplos. Cada triplo é composto por um sujeito, um predicado, e um objeto que são representados por URIs (RADIO *et al*., 2017, p. 165, tradução nossa).

O trabalho apresenta uma breve visão dos padrões que estão debaixo de guarda-chuvas disciplinares. Nas Ciências da Vida, dados e metadados são caracterizados por diversos eixos de variação. O enorme panorama de dados e metadados não estruturados reflete tais eixos. Um deles é a diversidade na escala dos fenômenos capturados por dados, o que desafia futuras tentativas de padronização de metadados descritores nesse campo. Apesar das especificidades entre seus subdomínios, os autores indicam a existência de padrões de metadados na biologia (EML) e na biodiversidade (*Darwin Core*, extensão do DC), ambos estruturados em árvore. Em algumas esferas de pesquisa quantitativa nas Ciências Sociais, conjuntos de dados possuem uma longa história de reúso, parte graças ao desenvolvimento e sucesso do *Data Documentation Initiative* (DDI), outro modelo estruturado em árvore (RADIO *et al*., 2017, p. 166-167, tradução nossa).

O DDI-*Lifecycle* vai além da especificação original do DDI-*Codebook* e assiste a gestão de dados científicos em todas as etapas de seu ciclo de vida, da coleta à análise, o arquivamento, e a descoberta. O DDI foi concebido para descrever dados de *surveys*, enquetes, e conjuntos de estatísticas (RADIO *et al*., 2017, p. 168, tradução nossa).

Para os autores, encontrar similaridades nas formas e estruturas que os dados científicos nas Ciências Humanas assumem é uma tarefa difícil. Os metadados para conjuntos de dados nas humanas são prescritos por padrões disponíveis em seus repositórios hospedeiros. Acreditam que o DC é o padrão mais presente na descrição de dados das humanas (RADIO *et al*., 2017, p. 169, tradução nossa).

Apontam para a dificuldade em qualificar qual tipo de atribuição caracteriza uma porção de dados como geoespacial. Entre os padrões mais notados à descrição de dados geoespaciais está a ISO 19115, que segue uma estrutura em árvore bem definida de forma a compartimentar aspectos dos dados (RADIO *et al*., 2017, p. 170, tradução nossa).

Descrever código criado durante o curso da pesquisa (ex.: modelo computacional) pode ser diferente se comparado à descrição de *software* que foi utilizado como parte da pesquisa (ex.: *software* comercial para análise de dados). No primeiro exemplo, metadados que dão créditos aos criadores do *software* serão necessários. Caso o objetivo seja sua preservação em longo prazo, os tipos de metadados requeridos serão provavelmente diferentes daqueles ao propósito da descoberta. Dados e *software* não existem em isolamento. Para possibilitar seu entendimento no contexto acadêmico, os relacionamentos entre eles precisam ser capturados. Esquemas de metadados à descrição de *software* estão em sua infância quando comparados àqueles que se destinam a dados. Geralmente, diferentes formas de descrever *software* não permitem interoperabilidade, tanto no significado do que está sendo descrito, quanto em termos de legibilidade por máquina (RADIO *et al*., 2017, p. 172-174, tradução nossa).

Concluem que independentemente do número de arquivos que o compõe, um conjunto de dados não pode ser considerado um todo ou entidade completa sem devida documentação, sendo essa condição necessária a sua contextualização. Afirmam que o valor real vem da combinação de conjuntos de dados e descritores (RADIO *et al*., 2017, p. 176, tradução nossa).

2.2.13 Yu (2017)

Yu (2017) acredita que os repositórios de acesso aberto têm significativamente alterado o panorama de arquivamento, armazenamento e curadoria de dados. Relata a obrigatoriedade da apresentação de planos de gestão de dados e sua disseminação por parte de agências de fomento e governos, o que tem forçado um aumento de serviços de gestão de dados em instituições e bibliotecas acadêmicas nos últimos anos. Serviços de dados científicos abrangem envolvimento ativo e apoio ao ciclo de vida dos dados, incluindo planejamento de gestão de dados; coleta, análise e armazenamento de dados; metadados e documentação de pesquisa; arquivamento, descoberta e compartilhamento de dados (YU, 2017, p. 783-784, tradução nossa).

O trabalho apresenta uma análise do atual nível dos serviços de dados científicos fornecidos por bibliotecas acadêmicas revisando a literatura recente. Seus objetivos foram: explorar o corrente estado dos serviços de dados científicos; contribuir com seu corpo de conhecimento; e informar a discussão sobre os desafios e oportunidades em seu desenvolvimento (YU, 2017, p. 784, tradução nossa).

Referindo-se ao fornecimento de serviços de dados científicos por bibliotecas acadêmicas, o autor exemplifica alguns dos desafios à prática identificados na literatura consultada, tais como comprometimento institucional, colaboração ampla no campus,

envolvimento do corpo docente, infraestrutura tecnológica, especialização limitada da(s) equipe(s) da(s) biblioteca(s), e carência de políticas e financiamento. Colaboração e centralização são reconhecidas tanto como desafios, quanto como oportunidades emergentes (YU, 2017, p. 786, tradução nossa).

De acordo com Yu (2017), a maioria dos estudos em sua revisão reconhece que bibliotecas acadêmicas estão em uma posição ideal para apoiar pesquisadores no enfrentamento dos desafios acerca da pesquisa intensiva em dados. Os dois tipos de serviços de dados científicos definidos na literatura consultada pelo autor são: informacional e técnico. O primeiro cobre uma variedade de serviços como consultoria em planos de gestão de dados e referência. O último implica proporcionar acesso a repositórios e sistemas de descoberta, preparar dados ou conjuntos de dados para serem depositados, e criar ou transformar metadados. A criação de políticas voltadas a serviços de dados científicos e a infraestrutura de gestão de dados são citadas como desafios e oportunidades (YU, 2017, p. 787, tradução nossa).

Para o autor, o armazenamento em nuvem de acesso aberto permitiu repositórios de dados se desenvolverem rapidamente. Em consequência do armazenamento em nuvem, obstáculos como o acesso a supercomputadores com poder de processamento e armazenamento de dados em nível de campus têm diminuído significativamente, o que sinaliza positivamente aos serviços de dados científicos (YU, 2017, p. 788, tradução nossa).

Yu (2017) cita que bibliotecas há muito se utilizam de metadados para facilitar a descoberta de informação, e que o arquivamento de dados fornece conservação em longo prazo e acesso a dados científicos. Ademais, informa que repositórios podem ser categorizados como de domínio (reconhecido por comunidade) e generalista, e que é importante que pesquisadores coloquem seus dados em dois ou mais repositórios ou arquivos (redundância de dados). Aponta que geralmente o escopo dos serviços inclui consultoria, treinamento, envolvimento ativo no planejamento da gestão de dados, orientação na gestão de dados durante a pesquisa, documentação e metadados de pesquisa, compartilhamento e curadoria de dados científicos (YU, 2017, p. 791-792, tradução nossa).

O autor identifica que muitas bibliotecas recomendam pesquisadores considerarem depositar seus dados científicos tanto em repositórios institucionais, quanto em repositórios mais apropriados para os tipos de dados gerados e seus potenciais usuários. Por conseguinte, instituições estão encorajando seus pesquisadores a depositarem seus dados em repositórios de acesso aberto, pois tal é exigido por algumas agências de fomento, assim como promove pesquisa aberta. Afirma que a maioria dos repositórios fornece suporte técnico na criação de metadados e documentação para assegurar futura descoberta, e que muitos também dão suporte à preservação de dados depositados ao migrar formatos de arquivos para evitar obsolescência digital (YU, 2017, p. 793, tradução nossa).

Conclui que muitas bibliotecas veem a prestação de serviços de dados científicos como uma forma de demonstrar seu valor no processo de criação do conhecimento. Entretanto, a falta de infraestrutura formal na gestão de dados e de políticas em muitas bibliotecas acadêmicas, a insuficiente preparação e treinamento de equipe(s), e o financiamento inadequado continuam como maiores obstáculos. Na era do *Big Data*, o fornecimento de serviços de dados científicos é um dos papéis principais de bibliotecas acadêmicas (YU, 2017, p. 793, tradução nossa).

Para Yu (2017), seu estudo de revisão revela que há demandas e oportunidades para explorar e investigar necessidades de gestão de enormes conjuntos de dados científicos produzidos pelo corpo docente; para repensar de forma a fornecer serviços e apoio ao ciclo de vida dos dados; para estrategicamente desenvolver políticas, infraestrutura e fluxo de trabalho aos serviços de dados científicos; para promover bibliotecários como parceiros nas áreas de gestão de conteúdo e acesso a ativos de pesquisa pertencentes ao corpo docente; e para proporcionar alerta ao corpo docente quanto a repositórios de dados de acesso aberto, encorajando assim efetiva gestão e preservação (YU, 2017, p. 793, tradução nossa).

3 PROCEDIMENTOS METODOLÓGICOS

A proposta de estudo e investigação aqui apresentada consiste em uma pesquisa aplicada, do ponto de vista de sua natureza; de método qualitativo, pelos meios de sua abordagem ao problema; exploratória e analítica, do ponto de vista dos objetivos; bibliográfica e documental, a partir dos procedimentos técnicos.

De acordo com Creswell (2013), a pesquisa qualitativa é uma abordagem para explorar e entender o significado que indivíduos ou grupos atribuem a um problema social ou humano. O pesquisador que se utiliza dessa abordagem geralmente levanta dados dos ambientes dos participantes, faz análises construídas a partir de temas específicos a genéricos, e faz interpretações sobre os significados dos dados. Logo, é um método indutivo que foca no significado individual e na importância da representação da complexidade de uma determinada situação (CRESWELL, 2013, p. 31, tradução nossa).

Creswell (2013) também afirma que, ao se planejar um estudo, o pesquisador precisa explicitar uma visão de mundo filosófica, compatível com o desenho ou modelo da pesquisa, que possua métodos específicos que traduzam essa abordagem em prática (CRESWELL, 2013, p. 33, tradução nossa).

3.1 Procedimentos do levantamento bibliográfico

Para o levantamento bibliográfico com vistas à concepção da revisão foram escolhidas quatro bases de dados de periódicos científicos. A escolha dessas bases se deve pela sua relevância dentro do campo da CI e em áreas inter e multidisciplinares na pesquisa científica mundial.

Nesse grupo, encontram-se bases especializadas (LISTA, LISA, ISTA) e a base multidisciplinar *Web of Science*. A *search string* utilizada foi ("*Research dat**" *AND* "*Metadata*"), considerando que todos os trabalhos indexados em bases internacionais são descritos por título, resumo e palavras-chave em Inglês. Inicialmente, a escolha dos termos a essa estratégia se deu pela própria natureza exploratória da pesquisa. Assim, termos foram utilizados a fim de recuperar textos que estivessem discutindo dados científicos e seus metadados (que se destinam à descrição dos dados), o que indiretamente leva à abordagem dos repositórios de dados e dos Princípios FAIR. Notou-se também que as palavras-chave escolhidas ao levantamento bibliográfico aparecem com significativa frequência durante leitura das diretrizes que compõem os FAIR. Em consequência disso, entende-se que tais termos possuem relativa relevância dentro do contexto e objetivo dessa investigação.

A pesquisa se limitou a recuperar artigos pelo uso dos termos supracitados no campo título (alto nível de relevância e especificidade), tendo como filtro de data publicações entre 2008 e 2018 (levantamento feito em abril de 2019). Outro critério de inclusão foi limitar a pesquisa a artigos revisados por pares, pois essa metodologia indica maior critério na avaliação desses estudos. Eliminaram-se estudos educacionais (ex.: competência informacional ou em dados), filosóficos, *surveys*, assim como aqueles que não discutiam os objetos aqui estudados.

Em levantamento inicial na base *Information Science & Technology Abstracts* (ISTA), com todos os critérios mencionados aplicados à estratégia de busca, recuperaram-se 8 artigos, onde apenas um era de acesso restrito. Portanto, 7 artigos atenderam aos critérios de seleção inicial.

Na base *Library, Information Science & Technology Abstracts* (LISTA), utilizando-se dos mesmos critérios aplicados na estratégia de busca anterior, recuperaram-se 11 artigos. Do número total, dois trabalhos eram de acesso restrito e sete foram encontrados também na ISTA. Dessa maneira, apenas 2 artigos atenderam aos critérios de seleção nessa etapa.

Ao iniciar a busca pela base *Library and Information Science Abstracts* (LISA) com o mesmo procedimento detalhado anteriormente (pesquisa no título, entre 2008 e 2018, revisado por pares), observou-se a necessidade de mudar a estratégia utilizada, pois ela não recuperou artigos. A *string* teve que ser modificada para ("*Research data*" *AND* "*Metadata*") e aplicada a todos os campos, menos ao texto completo (NOFT[18]). Também se utilizou de filtros como idioma (Inglês) e tipo de documento (artigo). Desse modo, recuperou-se um total de 41 artigos. Após a leitura de título, resumo e palavras-chave, chegou-se a 23 artigos, sendo que três deles eram de acesso restrito e cinco já haviam sido recuperados em buscas anteriores, o que resultou na seleção de 15 artigos.

Dessa forma, nas bases especializadas escolhidas se obteve um total de 24 artigos. Seguindo o planejamento do levantamento bibliográfico, fez-se então a busca na base multidisciplinar *Web of Science*. Aplicaram-se os mesmos critérios de busca, onde foram recuperados onze artigos, em que três tinham acesso restrito e outros três recuperados anteriormente. Ao final, obtiveram-se 5 artigos selecionados pelos critérios estabelecidos.

Com o processo de levantamento bibliográfico finalizado, obteve-se um total de 29 artigos, que passaram por outra avaliação (leitura de título, resumo, palavras-chave e texto completo). Após a avaliação de conteúdo, chegou-se a 13 artigos considerados relevantes à concepção de marco teórico pilar a essa investigação. A revisão da literatura foi apresentada seguindo dois critérios: primeiro, organização por ordem cronológica crescente de ano de publicação e; segundo, organização por ordem alfabética de autor(es).

[18] *No Full Text*.

3.2 Procedimentos da seleção dos objetos de estudo: repositórios

Os documentos analisados na pesquisa são de natureza digital, mais especificamente, conjuntos de dados científicos, ou seja, metadados e arquivos de dados observados por meio do acesso a repositórios de dados científicos de IES e agências de pesquisa sul-americanas recuperados pelo RE3DATA (*Datacite*), base de dados onde se indexam repositórios de dados e os descrevem com uso de metadados em *tags* como *description*, *repositoryContact*, *type*, *size*, entre outros. A Figura 4 apresenta o esquema de metadados da RE3DATA que serve à descrição de repositórios de dados científicos.

Figura 4 – Esquema de metadados da RE3DATA para descrição de repositórios de dados científicos.

```
<!--
  re3data.org Metadata Schema for the Description of Research Data Repositories. Version 3.0, December 2015. doi:10.2312/re3.008
 -->
<r3d:re3data xmlns:r3d="http://www.re3data.org/schema/3-0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.re3data.org/schema/3-0
 http://schema.re3data.org/3-0/re3dataV3-0.xsd">
  <r3d:repository>
    <r3d:identifiers>
      <r3d:re3data>r3d100000001</r3d:re3data>
      <r3d:doi>http://doi.org/10.17616/R34S33</r3d:doi>
    </r3d:identifiers>
    <r3d:repositoryName language="eng">Global Research Data Repository</r3d:repositoryName>
    <r3d:additionalName language="eng">GReDaR</r3d:additionalName>
    <!--  optional  -->
    <!--  multiple  -->
    <r3d:additionalName language="deu">Globales Forschungsdatenrepositorium</r3d:additionalName>
    <!--  optional  -->
    <!--  multiple  -->
    <r3d:repositoryUrl>http://www.globalresearchdatarepository.org</r3d:repositoryUrl>
    <r3d:repositoryIdentifier>
      <!--  optional  -->
      <!--  multiple  -->
      <r3d:repositoryIdentifierType>DOI</r3d:repositoryIdentifierType>
      <r3d:repositoryIdentifierValue>http://doi.org/0123456789/gredar</r3d:repositoryIdentifierValue>
    </r3d:repositoryIdentifier>
    <r3d:description language="eng">
      The Global Research Data Repository (GReDaR) is the place where researchers from all academic disciplines can put their research data. Research data deposited in
      GReDaR.
    </r3d:description>
    <!--  optional  -->
    <r3d:repositoryContact>info@greadar.org</r3d:repositoryContact>
    <r3d:type>disciplinary</r3d:type>
    <!--  multiple  -->
    <r3d:size updated="2012-11-23">10.000 datasets; 323 studies</r3d:size>
    <!--  optional  -->
    <r3d:startDate>2011-01-01</r3d:startDate>
    <!--  optional  -->
    <r3d:endDate>
```


Fonte: RE3DATA, 2020.

O RE3DATA.org é um registro global de repositórios de dados científicos que cobre repositórios de diferentes disciplinas acadêmicas. Apresenta repositórios para o arquivamento e acesso permanente de conjuntos de dados a pesquisadores, corpos financiadores, editores e instituições acadêmicas. O RE3DATA.org promove a cultura de compartilhamento, amplo acesso e melhor visibilidade de dados científicos (RE3DATA.ORG, 2019, *online*, tradução nossa).

De antemão, à luz de critérios científicos, decidiu-se que repositórios não indexados na base supracitada não fariam parte do universo investigado. Essa decisão foi tomada pela dificuldade da descoberta e acesso a repositórios não indexados em bases referenciais. A seguir, apresentam-se os processos decisórios quanto ao alcance e escolha dos objetos de estudo da investigação em questão.

3.2.1 Primeira etapa

Em um primeiro momento, quando da intenção de encontrar os objetos de estudo da então pretendida pesquisa, a estratégica de busca foi a exploração dos termos “repositório” e “dados científicos” na ferramenta de pesquisa do *Google*, pois se pretendia obter um primeiro contato com possíveis repositórios de dados científicos associados a instituições brasileiras.

Os resultados foram insatisfatórios devido à dificuldade de identificação desses repositórios entre as diversas páginas recuperadas pelo *Google*. Foi então que se obteve conhecimento da base RE3DATA e da possibilidade da recuperação de repositórios digitais de dados científicos por meio de *browsing*[19] entre as categorias: Assunto, Tipo de Conteúdo (ex.: imagem, código de fonte, entre outros) e País.

A partir daí, fez-se levantamento dos repositórios brasileiros na referida base, utilizando-se de estratégia de busca “por país” com objetivo de conhecer rapidamente os repositórios indexados. Após o primeiro contato, foi preciso estabelecer critérios de seleção para que se sustentasse rigores científicos.

Oito repositórios de dados científicos foram encontrados indexados como brasileiros na base RE3DATA, porém apenas a metade foi selecionada a partir dos critérios seguintes: a) repositórios geridos por grupos ou instituições brasileiras em ambientes controlados (repositórios institucionais, não bancos de dados e/ou arquivos pessoais de pesquisadores); b) possibilidade de busca de registros (conjuntos de dados depositados em repositórios) por *browsing* e; c) repositórios ativos e de acesso aberto.

Os repositórios selecionados na primeira etapa são apresentados na base RE3DATA como segue:

1) PPBio [20] *Data Repository* (Repositório de Dados de Levantamentos Biológicos);
2) IBICT *Dataverse Network* (Instituto Brasileiro de Informação em Ciência e Tecnologia *Dataverse Network*);
3) CEDAP *Research Data Repository – research data* (Centro de Documentação e Acervo Digital da Pesquisa – Dados de Pesquisa) e;
4) Base de Dados Científicos da Universidade Federal do Paraná (*Scientific Database of the Federal University of* Paraná).

[19] Técnica de busca exploratória onde se procura de unidade a unidade em um catálogo ou acervo, até que se encontre o desejado ou se esgote todas as opções.
[20] Programa de Pesquisa em Biodiversidade do Instituto Nacional de Pesquisa da Amazônia (INPA).

Feito a primeira análise dos repositórios recuperados, identificaram-se inconsistências dos *links* de acesso ao terceiro repositório supracitado, CEDAP, de responsabilidade da Universidade Federal do Rio Grande do Sul (UFRGS). Para além das dificuldades de acesso ao endereço indicado pelo *link* disponível, reconheceram-se inconsistências de acesso aos registros e seus conjuntos de dados a serem analisados. Após algumas tentativas, decidiu-se por excluir o repositório do estudo, limitando assim a exploração a 3 repositórios nacionais, o que foi considerado insuficiente para fins comparativos dentro da proposta de pesquisa.

3.2.2 Segunda etapa

A partir da avaliação dos objetivos do estudo em relação à quantidade de repositórios a serem investigados, decidiu-se por aumentar o escopo da investigação em nível continental. Desse modo, num novo acesso à base RE3DATA, foram encontrados 12 repositórios de dados divididos entre outros 4 países sul-americanos: Argentina, Chile, Colômbia e Peru. À vista dos critérios de seleção anteriormente estabelecidos, com exceção ao primeiro que agora se expande aos demais países do continente, selecionaram-se outros 5 repositórios de origem chilena, colombiana e peruana. Nenhum repositório argentino atendeu aos critérios estabelecidos.

Portanto, os repositórios selecionados na segunda etapa são:

5) *Repositorio de Datos de Investigación de la Universidad de Chile* (Chile);
6) CIAT[21] *Dataverse* (Colômbia);
7) *Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú* (Peru);
8) *Repositorio de datos del Ministerio de Educación del Perú* (Peru);
9) *Repositorio Institucional* USIL[22] (Peru).

A indexação de repositórios na base RE3DATA é de responsabilidade de instituições interessadas em tornar seus dados encontráveis, assim como sugere o primeiro princípio FAIR. Ela proporciona visibilidade na busca por repositórios e conjuntos de dados científicos, principalmente quando se parte do zero, na pretensão de explorar as variáveis dentro das inúmeras possibilidades de acesso na *Web*. A preocupação com anotação de esquemas de metadados e com a qualidade da indexação/disponibilização dos dados sobre esses repositórios em bases como a RE3DATA deve fazer parte do cotidiano da equipe

[21] Centro Internacional de Agricultura Tropical, membro do consórcio *Consultative Group on International Agricultural Research* (CGIAR).
[22] *Universidad San Ignacio de Loyola*.

gestora responsável por acervos de conjuntos de dados científicos, pois é também por serviços como esse que se pode garantir apropriada visibilidade.

Para que um repositório seja registrado e indexado no RE3DATA, ele precisa “ser gerido por uma entidade legal, como uma instituição sustentável (ex.: biblioteca, universidade); esclarecer condições de acesso aos dados e repositório assim como os termos de uso; e ter foco em dados científicos”. (RE3DATA, 2019, *online*, tradução nossa). Caso um repositório de dados científicos específico não esteja indexado na base RE3DATA, é possível sugeri-lo via aba *Suggest* localizada em seu endereço eletrônico.

3.3 Procedimentos da seleção dos objetos de estudo: conjuntos de dados

Considerou-se satisfatória a seleção dos 8 repositórios geridos por instituições localizadas na américa do sul, pois ela está dentro das expectativas da investigação que eram também possibilitar um quadro comparativo entre as variáveis estudadas.

O universo amostral da pesquisa contou então com 8 repositórios: 3 brasileiros, 1 chileno, 1 colombiano e 3 peruanos. Obteve-se a partir daí uma população (N) de 1.115 conjuntos de dados científicos. Esse valor é resultado da soma dos conjuntos identificados em cada repositório. Devido ao volume de conjuntos, tornou-se inviável sua coleta e análise de forma integral.

Dessa maneira, decidiu-se por utilizar uma técnica estatística chamada Amostragem Aleatória Estratificada. Sua escolha se deu pelo fato da mesma conseguir selecionar aleatoriamente uma quantidade de indivíduos (n) dentro de sua população (N), que por sua vez é estratificada em camadas não-lineares, ou seja, 8 repositórios distintos e com quantidades de conjuntos de dados variantes entre si. Não seria possível definir um número específico como amostra (n) da população (N = 1.115), nem quais indivíduos deveriam ser investigados, sem que se incorresse em *bias*[23]. Portanto, para que se chegasse ao número correspondente à amostra retirada de cada repositório, multiplicou-se separadamente o número de conjuntos de cada repositório {N1, N2, N3, N4, N5, N6, N7, N8} por cem por cento, dividindo-o por N (1.115).

$$N1 \times 100\% \div N \ = n1$$

O resultado foi a porcentagem retirada da população (N) que cada repositório teve que contribuir em conjuntos de dados ao valor total da amostra {n1 + n2 + n3 ... + n8 = n}. Ao

[23] Decisão enviesada tomada pelo investigador.

final, somando-se o número de conjuntos correspondente a cada porcentagem retirada, a partir do tamanho (peso) de cada repositório, obteve-se o valor da amostra (n) igual a 258 conjuntos de dados, que corresponde a 23% da população (N). A Tabela 1 apresenta os números resultantes dessa operação em ordem decrescente de número de conjuntos por repositório.

Tabela 1 – Amostragem Aleatória Estratificada aplicada às populações dos repositórios selecionados.

	Repositórios América do Sul	**População (N)**			**Amostra (n)**	
		N	%		n	% N
1	**PPBio (Brasil)**	403,0	36,14%		145,7	13,06%
2	***Repositorio Institucional* USIL (Peru)**	256,0	22,96%		58,8	5,27%
3	**CIAT *Dataverse* (Colômbia)**	189,0	16,95%		32,0	2,87%
4	**IBICT *Dataverse* Cariniana (Brasil)**	139,0	12,46%		17,3	1,55%
5	***Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú***	44,0	3,94%		1,7	0,16%
6	***Repositorio de datos del Ministerio de Educación del Perú***	44,0	3,94%		1,7	0,16%
7	**Base de Dados Científicos UFPR (Brasil)**	31,0	2,78%		0,9	0,08%
8	***Repositorio de Datos de Investigación de la Universidad de Chile***	9,0	0,80%		0,1	0,01%
	TOTAL	**1.115,0**	**100%**		**258,2**	**23,15%**

Fonte: Dados da pesquisa, 2020.

Torna-se imprescindível observar que o número de conjuntos de dados científicos depositados nos repositórios investigados pode ser indicador da maturidade de cada projeto. Teoricamente, esses números devem estar em constante expansão, pois a obrigatoriedade do depósito e acesso aberto aos dados unida à cultura da conscientização do compartilhamento fará com que o volume do arquivamento digital cresça exponencialmente, talvez até acompanhando o ritmo da criação e coleta dos dados. Entretanto, para que ocorra, será preciso grandes investimentos na manutenção dos produtos e serviços ofertados por parte das instituições interessadas, ou seja, recursos humanos especializados e infraestrutura atualizada.

Para que se obtivesse números inteiros, foi feito arredondamento dos valores das amostras de cada repositório. Entretanto, não foi possível fazê-lo no caso do *Repositorio de Datos de Investigación de la Universidad de Chile*, pois estatisticamente ele está próximo a zero, sendo assim eliminado da investigação.

Com o valor das amostras determinado, foi preciso sortear números entre 1 e o último número da população de cada repositório, como nos casos dos repositórios *Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú* (N6) e *Repositorio de datos del Ministerio de Educación del Perú* (N7), onde $N6\ e\ N7 = \{1, 2, 3, \ldots 44\}$. Nesses casos, ambos obtiveram 0,16% como amostra da população (N), porcentagem que corresponde a 1,7

conjunto, cujo arredondamento se iguala a 2, sendo essa a quantidade de conjuntos avaliados nesses repositórios. Desse modo, foram selecionados aleatoriamente os conjuntos de número $n6 = \{19, 33\}$ e $n7 = \{17, 32\}$. Observa-se que esses números não são fixos e não funcionam como identificação, pois correspondem às posições dos conjuntos no acervo digital, ao passo que caso sejam inseridos mais conjuntos aos repositórios, esses sofrerão deslocamento, recebendo posições diferentes.

Todos os números sorteados foram registrados. Para tal, utilizou-se da função "ALEATÓRIOENTRE" do *software Microsoft Excel*. Os conjuntos de dados sem arquivos foram desconsiderados à coleta e análise, sendo retirados do processo investigativo (valores na Tabela 1 correspondem a esse critério).

3.4 Procedimentos da coleta dos dados

Quanto à coleta dos dados que serviram aos resultados desse estudo, pode-se frisar que ela foi feita em três etapas realizadas entre os meses de agosto de 2019 a janeiro de 2020.

3.4.1 Primeira etapa

Os dados que serviram à primeira etapa da análise foram coletados nos *Websites* dos 3 repositórios brasileiros selecionados: PPBio *Data Repository*; IBICT *Dataverse Network* e; Base de Dados Científicos da Universidade Federal do Paraná. A coleta foi registrada em planilhas, salvas em um arquivo com extensão .xlsx (*Microsoft Excel Open XML Spreadsheet*), nas categorias (colunas): Número do Depósito (número não fixo nos repositórios); URL ou URI do Depósito; Identificador Persistente do Depósito; além do registro das extensões encontradas nos arquivos observados durante a exploração dos conjuntos.

As categorias Número do Depósito, URL ou URI e Identificador Persistente servem à identificação de cada conjunto de dados científicos. As categorias criadas com as extensões dos arquivos encontrados tinham o intuito de: 1) obter retorno estatístico por formato e extensão incidente; e 2) registrar, investigar e descrever cada um deles. Não se pretendeu fazer levantamento da quantidade de arquivos salvos em cada formato e extensão.

Em cada planilha também foi registrado o título do repositório, os padrões ou esquemas de metadados identificados nele, e o título do *software* de operacionalização. Também foram feitas anotações das especificidades de cada repositório analisado, as quais serviram à análise qualitativa dos dados.

Os dados coletados nessa fase podem ser conferidos na Tabela 3 do Capítulo 4 desse trabalho.

3.4.2 Segunda etapa

Essa etapa repetiu a anterior e se utilizou do mesmo modelo de planilha, todavia, com a coleta de dados dos outros 4 repositórios selecionados restantes à análise, que são: CIAT *Dataverse* (Colômbia); *Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú*; *Repositorio de datos del Ministerio de Educación del Perú*; *Repositorio Institucional* USIL (Peru).

Os dados coletados nessa fase podem ser conferidos na Tabela 4 do Capítulo 4 desse trabalho.

3.4.3 Terceira etapa

A terceira e última etapa da coleta dos dados se destacou pelo tipo de levantamento realizado. Nela, as planilhas criadas possuíam as mesmas categorias anteriormente citadas referentes à identificação dos conjuntos de dados científicos.

A diferença é que nessa fase foi registrada, na categoria intitulada *CSIRO Data Rating*, notas entre 0 e 5 obtidas por meio de avaliação dos conjuntos selecionados estatística e aleatoriamente na ferramenta FAIR intitulada *5-Star Data Rating Tool*, apresentada na subseção seguinte. Ao final da avaliação de todos os conjuntos, foi calculada a média aritmética ponderada para a obtenção de números que representassem seus respectivos repositórios, possibilitando assim um quadro comparativo entre as variáveis estudadas.

Os números obtidos e registrados na primeira fase de avaliação passaram por processo de checagem, onde os conjuntos foram reavaliados a fim de corrigir possíveis erros, assim como observar e coletar dados quanto a continuidade dos serviços dos repositórios estudados.

As médias ponderadas serviram como referência à análise de conteúdo dos grupos de conjuntos (metadados e arquivos de dados) próximos a elas, pois representam, em maior porcentagem, a realidade dos repositórios.

Por fim, os dados coletados nessa fase podem ser conferidos nas Tabelas 5 e 6, e no Gráfico 1, localizados no Capítulo 4 desse trabalho. O gráfico reúne, numa escala percentual entre 0 e 100, as médias e as maiores notas obtidas por meio da referida ferramenta, assim como os critérios FAIR (entre 0 e 15) atendidos por repositório por meio da análise de conteúdo à luz desses princípios. Desse modo, pode-se facilmente visualizar os resultados de natureza quantitativa quanto ao nível FAIR desses conjuntos e seus respectivos repositórios, assim como perceber de forma objetiva as discrepâncias entre as avaliações promovidas.

3.5 Procedimentos da análise dos dados

Finalizada a etapa de definição dos objetos de estudo e em posse dos dados coletados, partiu-se para a análise qualitativa de conteúdo, que contou com uma análise exploratória dos dados referentes ao objetivo específico C, onde foi possível identificar e descrever a natureza dos dados a partir dos formatos e extensões dos arquivos que os continham. Essa etapa exigiu a investigação dos formatos e extensões dos arquivos não usuais em plataformas especializadas para então defini-los. Dessa maneira, a análise foi compilada e apresentada como resultado parcial nos Quadros 5, 6, 7 e 8 do Capítulo 4.

Por conseguinte, para que se atingisse o resultado final dessa investigação que corresponde ao objetivo específico D, o qual busca avaliar e comparar o nível FAIR dos dados e seus repositórios, após a coleta das notas obtidas por avaliações feitas na ferramenta de autoavaliação *5-Star Data Rating Tool*, seguiu-se com a análise dos dados em duas etapas.

O intuito foi comparar os resultados encontrados perante o uso da ferramenta que possui caráter semiautomático com aqueles percebidos por avaliação humana por meio da análise de conteúdo. Ambas as avaliações (*computer-based* e *human-based*) foram descritas nos Quadros 9, 10, 11 e 12, respectivamente. Logo, os dois últimos quadros descrevem as avaliações feitas por interpretação humana à luz dos Princípios FAIR estudados. Assim, a análise de conteúdo se resumiu em análise textual de vocabulário técnico.

3.5.1 Primeira etapa: avaliação com o uso da *5-Star Data Rating Tool*

A análise nessa etapa contou com a seleção de um dos inúmeros conjuntos de dados avaliados por repositório, com nota média ou próximo a ela, para mais ou menos, sendo utilizado como amostra representativa do serviço responsável por seu depósito. Como informado anteriormente, a ferramenta *online 5-Star Data Rating Tool* da CSIRO possibilita a avaliação de depósitos de conjuntos de dados científicos com base nos Princípios FAIR.

O gestor ou depositante interessado em avaliar – de forma semiautomática – a conformidade desses conjuntos e repositórios com os Princípios FAIR pode acessar a ferramenta e responder um questionário, que ao final, retorna resultado quantitativo como resposta ao nível FAIR dos objetos avaliados. Assim como o nome da ferramenta, o número obtido faz alusão às 5 Estrelas para Dados Abertos de Berners-Lee (*FIVESTARDATA*, 2019, *online*), pois é gerado numa escala entre 0 e 5 (estrelas).

A *5-Star Data Rating Tool* fornece um esquema de classificação que usa autoavaliação em relação aos atributos sociais, técnicos e informacionais dos dados. Esta ferramenta fornece implementações dos Princípios de dados FAIR da *FORCE 11*. O esquema das 5 Estrelas visa ajudar usuários a entender a maturidade de alguns dados ou serviços (CSIRO, 2017, *online*, tradução nossa).

Como visto, os Princípios FAIR possuem 15 critérios que devem ser atendidos dentro do processo de adaptação FAIR aos dados e metadados arquivados em seus repositórios. Todavia, a ferramenta faz modificações que se destoam minimamente dos Princípios FAIR. Em 14 questões, cada uma com opções entre 3 e 7 respostas distintas, 3 dos critérios FAIR (F4, A1.1 e A1.2) não estão contidos em sua proposta avaliativa – e não se sabe o porquê, assim como também não se possui conhecimento das métricas que levam aos resultados numéricos.

Embora não se tenha informação quanto ao sistema métrico da ferramenta, pode-se fazer uma conta básica na qual 5 (estrelas) divididas por 14 (questões) resulta o valor de 0,357[...] por questão. Caso as questões que não envolvem os Princípios FAIR fossem anuladas, obter-se-ia uma dízima periódica de 0,4545 [...] por questão – acreditando que o peso entre elas fosse o mesmo. O valor por questão se divide então entre as opções de resposta, onde a primeira tem pontuação mínima ou nula, e a última, máxima. O Quadro 4 evidencia o observado, assim como a correlação, em algumas proposições, entre as 5 Estrelas para Dados Abertos e os Princípios FAIR, pois um está contido no outro de forma indireta. Todavia, o quadro disponibilizado pela ferramenta se difere minimamente das questões dispostas no processo avaliativo (ex.: “Questão 2. Publicados” possui mais 4 opções de resposta).

Quadro 4 – Esquema de avaliação da *5-Star Data Rating Tool*.

Palavra-chave	Conjuntos de dados devem ser ou estar...	Princípios FAIR Correspondentes	Níveis	Faceta correspondente às 5 Estrelas para Dados Abertos Conectados
Publicados	… destinados a serem acessíveis a usuários além do criador ou dono	N/A	a. Sem acesso externo b. Acesso externo, protocolo não *Web* (ex.: distribuição de mídia física) c. Publicado via *Web*	c. ★
Hospedados	... disponíveis na *Web*	A1	a. Não está na *Web* b. Arquivos em servidor *Web* c. Repositório com interface *Web* d Serviço *Web* – API local e. Serviço *Web RESTful*[24] - *OpenAPI/Swagger* f. API *Web* padrão (SPARQL, OGC WMS/WFS/WCS/SOS/WPS, ...)	b. ★ c. ★ d. ★ e. ★ f. ★
Tratados	… fornecidos com compromisso que esses dados estarão disponíveis por longo prazo	A2	a. despejo único, sem compromisso contínuo b. Melhor esforço c. Repositório institucional d. Repositório certificado	
Atualizados, preservados	… parte de um programa ou série regular de coleta de dados, com claras disposições de manutenção e cronograma de atualização	N/A	a. Conjunto de dados único b. Parte de série de dados, atualização irregular/ocasional c. Parte de série de dados com atualizações regulares	
Licenciados	… claramente licenciados, para que as condições de reuso estejam	R1.1	a. Sem licença b. Licença descrita em texto	b. ★ c. ★

[24] REST (*REpresentational State Transfer*)

	disponíveis e expressas inequivocamente		c. Licença padrão (ex.: *Creative Commons*)	
Citáveis	… denotados usando um identificador persistente estável	F1	a. Não citável b. Identificador local (pode mudar) c. Identificador *Web* (URL ou consulta transitória) d. Identificador *Web* persistente (PURL, DOI, handle, ARK, etc.)	c. ★ d. ★★★★
Descritos	… descritos e marcados com metadados formais de conteúdo	R1, F2, F3	a. Sem metadados b. Descrição de texto (resumo) e palavras-chave c. Metadados básicos (ex.: *Dublin Core*) d. Metadados especializados (ex.: *Darwin Core*, ISO 19115, perfil de dados científicos do schema.org) e. Metadados ricos utilizando vocabulários RDF (ex.: DCAT, ADMS, PROV, GeoDCAT, OMV, VoID)	c. ★★ d. ★★ e. ★★
Encontráveis	… indexados em um sistema bem conhecido (pode ser de propósito geral ou específico de comunidade)	R1, F2, F3	a. Não indexado b. Indexado em um local, catálogo organizacional c. Metadados coletados ou inseridos em uma comunidade (ex.: *Research Data Australia*, RE3DATA) ou catálogo jurisdicional d. Visível em índices de uso geral (*Google, Bing*) e. Altamente classificado em índices de uso geral	d. ★★ e. ★★
Carregáveis	… representados usando um formato (pré-requisito: registro de formato de dados) comum ou endossado por comunidade (ex.: padrão) Obs.: formatos de páginas como .doc e .pdf que visam colocar pixels em uma página para consumo humano não contam! Dados fornecidos nesses formatos não podem ser carregados por aplicativos de processamento de dados padrão.	I1	a. Formato de arquivo customizado b. Um formato de dados padrão, denotado por um tipo MIME (CSV, JSON, XML, netCDF, etc.) c. Escolha entre múltiplos formatos padrões	a. ★★ b. ★★★ c. ★★★
Usáveis	… estruturados usando um esquema endossado por comunidade (padrão?) detectável ou modo de dados (pré-requisito: registro de modelos, esquemas, ontologias de dados)	I2, R1.3	a. Esquema implícito, não formalizado b. Esquema explícito, formalizado em DDL, XSD, pacote de dados, RDFS/OWL, JSON-*Schema* ou similar c. Esquema de comunidade, disponível de uma localização (padrão)	b. ★★ c. ★★★
Compreensíveis	… suportados com definições inequívocas para todos os elementos internos (ex.: definições de colunas, unidades de medidas), por meio de *links* a definições acessíveis (pré-requisito: registro de vocabulários de unidades de medida, quantidades, propriedades observáveis)	I2	a. Etiquetas de campo local b. Etiquetas de campo ligadas a explicações de texto c. Etiquetas padrões (ex.: CF *Conventions*, UCUM *units*) d. Alguns nomes de campo vinculados a vocabulários padrões geridos externamente e. Todos os nomes de campo vinculados a vocabulários padrões geridos externamente	
Conectados, vinculados	… vinculados a outros dados usando identificadores externos (ex.: URIs), potencialmente rastreáveis	I3	a. Sem *links* b. *Links* de acesso interno a partir de um catálogo ou página de destino c. *Links* de acesso externo a dados relacionados	b. ★ c. ★★★★★
Avaliáveis	… acompanhados por, ou conectados a uma avaliação da qualidade dos dados e descrição da origem e fluxo de trabalho que produziu os dados	R1.2	a. Sem informação de qualidade ou origem b. Declaração de origem em texto c. Rastreio formal de proveniência (W3C PROV-O ou similar)	
Confiáveis	… acompanhados por, ou conectados a informações sobre como os dados foram usados, por quem, e quantas vezes	N/A	a. Nenhuma informação sobre uso b. Estatística de uso disponível c. Claramente endossado por estrutura ou organização conceituada	

Fonte: CSIRO, 2017, *online*, tradução nossa.

Como observado no Quadro 4, a linha que se refere a dados publicados não está explicitamente contida nos Princípios FAIR, mas corresponde à primeira estrela das 5 Estrelas para Dados Abertos. Entretanto, de forma indireta e óbvia, para que sejam encontrados e acessados, os dados precisam estar publicados na *Web*. Quanto às linhas que se referem a dados atualizados/preservados e confiáveis, o quadro deixa claro que não há relação com os FAIR, nem com as 5 estrelas.

A fim de satisfazer a natureza dessa investigação, os resultados numéricos das avaliações foram apresentados em texto (respostas dadas) nos Quadros 9 e 10 do próximo capítulo, seguidos das percepções a partir dos processos de análise exploratória qualitativa.

3.5.2 Segunda etapa: análise de conteúdo baseada nos Princípios FAIR

A última etapa de avaliação contou com a percepção dos resultados obtidos pela ferramenta anteriormente citada, pois o processo de resposta ao questionário permitiu uma aproximação com a realidade dos repositórios e seus conjuntos de dados científicos, sendo esse o primeiro contato que se teve com esses objetos à luz dos Princípios FAIR. A partir disso, foi feita a releitura dos princípios em sua íntegra, e então a conferência de cada declaração do padrão com os objetos analisados, a fim de averiguar e expressar a sua (in)compatibilidade com o proposto. Assim, ao final dessa análise, foram construídos os Quadros 11 e 12, onde se percebem os critérios atendidos por repositório, seguidos dos motivos pelos quais o foram. Para além dos quadros, foram feitas observações que complementam e finalizam o processo dessa investigação.

3.6 Observações ao processo metodológico

Passada a fase de seleção dos repositórios investigados, obteve-se conhecimento da existência de uma base que indexa metadados referentes a conjuntos de dados científicos arquivados em repositórios de instituições de pesquisa do estado de São Paulo. Trata-se do Metabuscador de Dados de Pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Essa base, operacionalizada pelo *DSpace*, contém metadados que fazem referências às coleções de dados científicos das seguintes instituições: UFSCAR, UNIFESP, USP, UNESP, ITA, UFABC, UNICAMP, e EMBRAPA. Apesar da fase de seleção de repositórios ter sido finalizada, fez-se um exercício estatístico utilizando os valores encontrados anteriormente, somando-os aos números de conjuntos de dados encontrados nos repositórios paulistanos. O resultado apresentado na Tabela 2 se tornaria insatisfatório à investigação, pois quase todos os repositórios das instituições supracitadas foram estatisticamente eliminados, com exceção da EMBRAPA.

Tabela 2 – Amostragem Aleatória Estratificada aplicada às populações com adição do estado de São Paulo, Brasil.

	Repositórios América do Sul	População (N)			Amostra (n)	
		N	%		n	% N
1	**PPBio**	403,0	28,42%		114,5	8,08%
2	**EMBRAPA**	270,0	19,04%		51,4	3,63%
3	***Repositorio Institucional* USIL (Peru)**	256,0	18,05%		46,2	3,26%
4	**CIAT *Dataverse* (Colômbia)**	189,0	13,33%		25,2	1,78%
5	**IBICT *Dataverse* Cariniana**	139,0	9,80%		13,6	0,96%
6	***Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú***	44,0	3,10%		1,4	0,10%
7	***Repositorio de datos del Ministerio de Educación del Perú***	44,0	3,10%		1,4	0,10%
8	**Base de Dados Científicos UFPR**	31,0	2,19%		0,7	0,05%
9	**USP**	15,0	1,06%		0,2	0,01%
10	***Repositorio de Datos de Investigación de la Universidad de Chile***	9,0	0,80%		0,1	0,01%
11	**UFSCAR**	6.0	0,42%		0,0	0,00%
12	**UNICAMP**	5.0	0,35%		0,0	0,00%
13	**UNESP**	3.0	0,21%		0,0	0,00%
14	**UFABC**	3.0	0,21%		0,0	0,00%
15	**UNIFESP**	1,0	0,07%		0,0	0,00%
16	**ITA**	0,0	0,00%		0,0	0,00%
	TOTAL	**1.418,0**	**100%**		**254,7**	**17,95%**

Fonte: Dados da pesquisa, 2020.

Embora a Base de Imagens de Sintomas de Doenças de Plantas (PDDB) do repositório Digipathos - EMBRAPA tenha sido estatisticamente percebida como apta a ser investigada na presente pesquisa, há uma certa limitação com relação a sua natureza. Pode-se dizer que seus conjuntos de dados possuem formato homogêneo, se diferenciando apenas entre as extensões dos arquivos de imagens salvos. Sendo assim, o objetivo específico C "identificar e descrever os formatos [...]" não necessitaria de esforços para ser alcançado, o que facilitaria, porém pouco contribuiria com o processo investigativo.

Outra eventualidade que deve ser relatada e que diz respeito ao processo metodológico envolve as avaliações feitas com o uso da *5-Star Data Rating Tool* na fase da coleta de dados. Como informado anteriormente, a coleta foi feita em duas fases, uma de avaliação e outra de reavaliação dos conjuntos de dados selecionados arquivados em seus respectivos repositórios. Pois que, durante o período de reavaliação, tentou-se por vezes obter acesso aos conjuntos arquivados no repositório do IBICT, sem sucesso.

Não foi possível encontrá-los via endereço eletrônico registrado na primeira fase da coleta, nem por meio do RE3DATA ou *Google*, tampouco pelo próprio *Dataverse*. Acredita-

se que o repositório estava em fase de testes e foi retirado do ar por tempo indeterminado. Não se encontrou qualquer informação a esse respeito no *Website* do IBICT.

Dessa forma, sem a possibilidade de reavaliar o repositório e seus conjuntos, e por ele ir de encontro ao critério de seleção “c) repositórios ativos e de acesso aberto”, foi preciso retirá-lo da investigação. Decidiu-se por manter os dados da primeira fase de coleta, pois esses não interferem no resultado final da investigação, mas servem como complemento.

Como se sabe, a RE3DATA não possui todos os repositórios de dados científicos existentes na América do Sul indexados em sua base, uma vez que tal responsabilidade e decisão partem do interesse de cada instituição. Ainda, a literatura aponta para a base chamada *OpenDOAR*, onde é possível indexar e/ou recuperar repositórios de acesso aberto, não se limitando a dados científicos. Em uma breve busca por país na base *OpenDOAR*, somente foi possível encontrar o repositório de dados do IBICT (fora do ar) e o repositório institucional da UFPR em sua lista de repositórios brasileiros indexados. Dos demais repositórios sul-americanos recuperados via RE3DATA, apenas o da USIL (Peru) foi percebido na *OpenDOAR*, pois assim como a UFPR, seu repositório gerencia e preserva documentos bibliográficos e dados em um mesmo ambiente (*DSpace*).

4 ANÁLISE DOS RESULTADOS

Com o primeiro objetivo específico da investigação alcançado ao se levantar os repositórios de dados científicos existentes no continente sul-americano, partiu-se então para a análise dos dados coletados.

A análise desses dados foi executada em duas etapas: a primeira, que corresponde ao objetivo específico C, identifica e descreve os formatos e extensões dos arquivos que compõem os conjuntos de dados; e a segunda, que se refere ao objetivo específico D, verifica e compara a conformidade dos repositórios e seus conjuntos de dados científicos com os Princípios FAIR.

4.1 Análise exploratória dos formatos e extensões dos arquivos avaliados

4.1.1 Repositórios brasileiros

De forma a facilitar a visualização dos dados coletados e analisados, construíram-se duas tabelas onde se pode identificar as extensões dos arquivos depositados em cada conjunto de dados científicos investigado, assim como a sua devida incidência. Ao conjunto que apresentou mais de um arquivo com a mesma extensão foi contabilizado apenas uma ocorrência em tabela, desse modo, os números registrados são referentes à incidência de cada extensão por depósito. A Tabela 3 apresenta as extensões encontradas, assim como a estatística resultante de sua incidência nos repositórios brasileiros em questão.

Tabela 3 – Extensões dos arquivos em conjuntos de dados científicos nos repositórios brasileiros.

Repositórios brasileiros	.txt	.doc	.docx	.pdf	.csv	.xlsx	.gpx	.gdb	.data	.zip	.rar	.xml	.rdf	**Total por repositório**
PPBio *Data Repository*	119	2	N/A	12	26	2	1	1	2	N/A	N/A	146	N/A	311
	38,26%	0,64%	N/A	3,86%	8,36%	0,64%	0,32%	0,32%	0,64%	N/A	N/A	46,95%	N/A	100%
IBICT *Dataverse Network*	N/A	1	4	13	N/A	N/A	N/A	N/A	N/A	1	1	N/A	N/A	20
	N/A	5,00%	20,00%	65,00%	N/A	N/A	N/A	N/A	N/A	5,00%	5,00%	N/A	N/A	100%
B. de Dados Científicos da UFPR	N/A	N/A	N/A	N/A	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	1	2
	N/A	N/A	N/A	N/A	N/A	50%	N/A	N/A	N/A	N/A	N/A	N/A	50%	100%
Total de extensões por conjunto	119	3	4	25	26	3	1	1	2	1	1	146	1	333
	35,74%	0,90%	1,20%	7,51%	7,81%	0,90%	0,30%	0,30%	0,60%	0,30%	0,30%	43,84%	0,30%	100,00%

Fonte: Dados da pesquisa, 2020.

A maioria (8 em 13) dos formatos ou extensões encontrados são parte da rotina de pessoas que utilizam computadores com certa frequência, como os arquivos de natureza textual (.txt, .doc, .docx, e .pdf), visual alfanumérica (.csv e .xlsx), ou aqueles que fazem compactação de outros arquivos (.zip e .rar). Não foi objetivo desse trabalho investigar o conteúdo de arquivos compactados encontrados durante o levantamento.

As demais extensões apresentadas na Tabela 3 precisaram ser investigadas para que se pudesse identificar e descrever a sua natureza. Dessa forma, o Quadro 5 apresenta a definição de cada extensão para melhor entendimento dos conteúdos salvos nos arquivos analisados.

Quadro 5 – Definições de extensões não usuais.

Extensão	Definição
.gpx	Um arquivo GPX é um arquivo de dados de GPS salvos no formato GPS *Exchange*, o qual é um padrão aberto que pode ser livremente usado por programas de GPS. Contém dados de localização representados por longitude e latitude que pode incluir pontos de notificação, rotas e trajetos. Arquivos GPX são salvos em formato XML, de forma que permite dados de GPS serem mais facilmente importados e lidos por múltiplos programas e serviços *Web*.
.gdb	Um arquivo GDB é um arquivo de base de dados criado por *MapSource*, um aplicativo editor de rotas GPS e de planejamento de viagens. Contém pontos de notificação, rotas, e trajetos que podem ser transferidos a um dispositivo de navegação Garmin. Arquivos GDB são similares aos universalmente transferíveis arquivos .GPX.
.data	Um arquivo DATA é um arquivo de dados usado por *Analysis Studio*, um programa de mineração de dados e análise estatística. Contém dados minerados em um texto simples, formato tabelar delimitado, incluindo um cabeçalho de arquivo do *Analysis Studio*. Arquivos DATA são comumente usados para armazenar dados para análise de dados offline quando não conectado a um servidor *Analysis Studio*, contudo pode também ser usado em modo *online*.
.xml	Um arquivo XML é um arquivo de dados XML. É formatado quase como um documento .HTML, mas usa *tags* personalizadas para definir objetos e os dados dentro de cada objeto. Arquivos XML podem ser pensados como uma base de dados baseada em texto.
.rdf	Um arquivo RDF é um documento escrito em linguagem RDF, que é usada para representar informação sobre recursos na *Web*. Contém informação sobre um *Website* em um formato estruturado chamado metadado. Arquivos RDF podem incluir um mapa de site, um registro de atualizações, descrições de páginas, e palavras-chave.

Fonte: *SHARPENED PRODUCTIONS*, 2020, *online*, tradução nossa.

Como observado na Tabela 3, as extensões de arquivos com maior frequência no repositório do PPBio estão em formato de texto (.txt) com 38,26%, e em formato XML com incidência em todos os conjuntos investigados (146). Arquivos .xml guardam os dados, metadados e suas *tags* anotadas referentes aos conjuntos depositados no repositório, o que facilita a importação de registros a partir de seu uso (interoperabilidade). O repositório do PPBio foi aquele com o maior volume de conjuntos de dados científicos investigado (403), e conta basicamente com um acervo textual e numérico tabelar. Os conjuntos de dados associados ao PPBio estão disponibilizados na infraestrutura tecnológica provida pela iniciativa *DataONE*, comunidade internacional que compartilha acervos de dados científicos nos campos da Biologia e das Ciências da Natureza. O *DataONE* utiliza a aplicação *Metacat* (escrita em *Java* e mantida por servidor *Apache Tomcat*) que gerencia acervos das áreas supracitadas.

Metacat é um catálogo flexível de metadados de código-fonte aberto e um repositório de dados com foco em dados científicos, particularmente advindos da Ecologia e

das Ciências do Meio Ambiente. Ele adota a linguagem XML como uma sintaxe comum para representar o vasto número de padrões de conteúdo de metadados que são relevantes à ecologia e outras ciências. Ademais, o *Metacat* é uma base de dados XML genérica que permite armazenamento, consulta, e recuperação de documentos XML arbitrários sem conhecimento prévio do esquema XML. Está sendo utilizado extensivamente ao redor do mundo para gerir dados do meio ambiente (*DATAONE*, 2020, *online*, tradução nossa).

No que diz respeito ao processo de inserção e descrição de dados, o usuário desse sistema pode inserir metadados por meio do *Morpho*, *software* editor de metadados também utilizado pelo *DataONE* em assistência à gestão e curadoria de acervos digitais no campo da Ecologia.

Morpho é um programa que pode ser usado para inserir metadados, que são então armazenados em um arquivo que obedece à especificação da *Ecological Metadata Language* (EML). Informação sobre pessoas, lugares, métodos de pesquisa, e atributos de dados estão entre os metadados coletados. Dados podem ser armazenados com os metadados no mesmo arquivo. Ele permite o usuário criar um catálogo local de dados e metadados que podem ser consultados, editados e visualizados. O *Morpho* também faz interface com o servidor *Metacat* chamado *Knowledge Network for Biocomplexity*[25] (KNB), que permite cientistas enviarem, baixarem, armazenarem, consultarem e visualizarem dados e metadados públicos (*DATAONE*, 2020, *online*, tradução nossa).

Adiante, o repositório do IBICT teve incidência maior de arquivos .pdf, que são arquivos em formato de texto protegidos de qualquer alteração, como devem ser publicações científicas, trabalhos acadêmicos e documentos oficiais. São esses os casos encontrados nos arquivos acessados. Cabe a crítica ao arquivamento de trabalhos completos em repositórios de dados científicos que não descreve o endereço onde esses foram primariamente publicados na *Web* (ausência de URI ou identificador persistente que os localize, por exemplo).

Por conseguinte, não faz sentido se utilizar de repositórios de dados científicos para depositar apenas arquivos de textos referentes à comunicação científica, duplicando algumas vezes os arquivos sem os interconectá-los. Exemplo disso seria uma instituição que possui um repositório direcionado à gestão e curadoria de documentos bibliográficos e também um repositório de dados científicos, onde uma mesma publicação é depositada em ambos repositórios, sendo que o segundo também recebe – ou deveria receber – o depósito dos dados relativos a ela. Sem a devida indicação de correlação e coexistência entre os depósitos por meio de *link*, haverá duplicação despercebida por parte de seus gestores.

[25] Rede de Conhecimento à Biocomplexidade (RCB).

Até onde se pôde apurar antes de sair do ar, o IBICT manteve seu repositório de dados científicos no *software Dataverse*, uma comunidade que compartilha repositórios de todo o globo nas diversas áreas do conhecimento humano e é um projeto de responsabilidade da Universidade de *Harvard* (*Cambridge, Massachusetts*).

Dataverse é um aplicativo *Web* de código aberto para compartilhar, preservar, citar, explorar, e analisar dados científicos. Facilita o processo de depósito, publicação e recuperação dos dados, tornando-os disponíveis a terceiros, e permite replicar trabalhos mais facilmente. Pesquisadores, periódicos, autores de dados, editores, distribuidores de dados, e instituições afiliadas recebem crédito acadêmico e visibilidade na *Web*. Um repositório *Dataverse* é uma instalação deste *software*, que então hospeda múltiplos arquivos virtuais chamados *Dataverses*. Cada *dataverse* contém conjuntos de dados, e cada conjunto contém metadados descritivos e arquivos de dados (incluindo documentação e código que acompanham os dados). Como um método de organização, os *dataverses* podem também conter outros *dataverses* (*DATAVERSE*, 2020, *online*, tradução nossa).

Em ambos os casos de comunidades compartilhadas, o poder de inclusão de acervos em suas bases é dado a instituições de pesquisa devidamente cadastradas. Para servir ao propósito da interoperabilidade, o *Dataverse* possibilita a exportação de metadados nos formatos *JavaScript Object Notation* (JSON, nativo do *Dataverse*), *Dublin Core*, DDI, DDI HTML *Codebook*, *DataCite* 4, OAI_ORE, *OpenAIRE* e *Schema.org* JSON-LD.

O último repositório analisado nessa etapa foi a Base de Dados Científicos da Universidade Federal do Paraná (UFPR). Observou-se que ela faz parte do Repositório Digital Institucional da UFPR, onde se encontram depósitos de documentos variados produzidos pela universidade, tais como Teses e Dissertações, Trabalhos de Especialização, Trabalhos de Graduação, entre outros. Ambos os repositórios compartilham o mesmo ambiente e são geridos por meio do mesmo *software* de código-fonte aberto chamado *DSpace*. O "*DSpace* foi desenvolvido para possibilitar a criação de repositórios digitais com funções de armazenamento, gerenciamento, preservação e visibilidade da produção intelectual, permitindo sua adoção por outras instituições em forma consorciada federada" (IBICT, 2019, *online*).

Como o *DSpace* não se destina à criação de comunidades de compartilhamento de repositórios e conjuntos de dados científicos entre diferentes instituições como o *DataONE* e o *Dataverse*, a sua utilização reduz as possibilidades de acesso a dados científicos em potencial dentro de um cenário de consulta exploratória. Apesar do *DSpace* não ter sido originalmente projetado para lidar com dados científicos, o mesmo pode ser customizado para tal fim, se assim for necessário.

Além do repositório de dados científicos da UFPR possuir o menor acervo dentre os repositórios brasileiros qualificados à investigação, ele também se confunde com o

repositório institucional durante navegação. Pôde-se verificar que 24 dos seus 30 conjuntos de dados depositados possuem arquivos .rdf que contêm *tags* e textos referentes ao uso da licença *Creative Commons*.

4.1.2 Demais repositórios sul-americanos

Os critérios apresentados na primeira etapa de análise dos dados foram replicados aqui, de forma a padronizar as ações para que se pudesse comparar os resultados obtidos. Assim, construiu-se a Tabela 4 nos mesmos moldes da anterior, sendo ela a referência das análises dessa subseção.

Tabela 4 – Extensões dos arquivos em conjuntos de dados científicos nos demais repositórios sul-americanos.

Demais Repositórios sul-americanos	.txt	.docx	.pdf	.csv	.xls	.xlsx	.xlt	.pptx	.jpg	.png	.tif	.tab	.dta	.zip	.sav	.do	Total por Repos.
CIAT (Colômbia)	1	3	9	1	19	2	1	1	1	1	1	23	2	7	N/A	2	74
	1,35%	4,05%	12,16%	1,35%	25,68%	2,70%	1,35%	1,35%	1,35%	1,35%	1,35%	31,08%	2,70%	9,46%	N/A	2,70%	100%
P. *Datos Abiertos* PUC (Peru)	N/A	N/A	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2	N/A	N/A	N/A	N/A	4
	N/A	N/A	50,00%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	50,00%	N/A	N/A	N/A	N/A	100%
R. *Datos* MEC (Peru)	N/A	N/A	2	2	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5
	N/A	N/A	40%	40%	N/A	20%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100%
USIL (Peru)	N/A	N/A	2	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1	1	54	N/A	59
	N/A	N/A	3,39%	N/A	1,69%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1,69%	1,69%	91,53%	N/A	100%
Total formatos e extensões	1	3	15	3	20	3	1	1	1	1	1	25	3	8	54	2	142
	0,70%	2,11%	10,56%	2,11%	14,08%	2,11%	0,70%	0,70%	0,70%	0,70%	0,70%	17,61%	2,11%	5,63%	38,03%	1,41%	100%

Fonte: Dados da pesquisa, 2020.

Torna-se relevante observar que entre os demais países investigados, apenas a Colômbia e o Peru possuem repositórios qualificáveis a este estudo. Também se deve notar que o Peru detém 3 dos 6 repositórios investigados em toda a América do Sul. Em contrapartida, o Brasil se responsabiliza por dois repositórios que juntos detêm o maior volume de dados científicos arquivados.

Entende-se que o volume não está diretamente relacionado à qualidade da gestão desses repositórios, tampouco diz respeito à situação financeira de uma instituição, uma vez que dados não são adquiridos por meio de compra, como livros e outros materiais. Entretanto, o volume de dados em um repositório pode indicar determinada maturidade de seus serviços.

A informação de que o repositório do PPBio (Brasil) tem o maior volume de conjuntos de dados compartilhados pode indicar maior envolvimento dos pesquisadores por

meio da conscientização do compartilhamento de seus dados, ou compartilhamento obrigatório via política de financiamento (como exigido dos pesquisadores financiados pela FAPESP).

Entre os conjuntos investigados depositados no repositório da *Universidad San Ignacio de Loyola*, notou-se a incidência majoritária da extensão .sav, que é descrita no Quadro 6. Outras extensões não usuais encontradas também estão presentes no quadro mencionado.

Quadro 6 – Outras definições de extensões não usuais.

Extensão	Definição
.tif	Um arquivo TIF é um arquivo de imagem salvo em um formato gráfico de alta qualidade. É usualmente utilizado para armazenar imagens com muitas cores, tipicamente fotos digitais, e inclui suporte para camadas e múltiplas páginas.
.tab	Arquivo de texto que contém uma tabela de dados na qual colunas são separadas por abas; pode ser importado pela maioria dos programas de planilha, que formatarão os dados em células.
.dta	Arquivos que contêm a extensão .dta são comumente associados com uma variedade de aplicativos, para uma série de formatos de arquivos de dados. Os arquivos DTA mais comuns são normalmente armazenados em formato binário ou textual. O programa *Turbo Pascal* usa o formato de arquivo DTA para os arquivos de dados referenciados pelo aplicativo. O *software Stata* [programa de estatística] também usa o formato de arquivo DTA para arquivos de conjuntos de dados salvos.
.sav	Arquivo de dados criado pelo *Statistical Package for the Social Sciences*[26] (SPSS), um aplicativo usado para análise estatística; salvo em um formato binário proprietário e contém um conjunto de dados assim como um dicionário que o descreve; salva dados por "casos" (linhas) e "variáveis" (colunas).
.do	Um arquivo DO é um programa *Java* baseado na *Web* executado por um servidor *Web* que suporta Java, como *Tomcat* ou IBM *WebSphere*. É tipicamente mapeado ao componente *Controller* do *framework Struts*, que processa o arquivo. Arquivos DO são utilizados para gerar páginas *Web* dinâmicas.
.png	Um arquivo PNG é um arquivo de imagem armazenado no formato *Portable Network Graphic* (PNG). Contém um *bitmap* de cores indexadas e usa compactação sem perdas, similar a um arquivo .gif, mas sem limitações de direitos autorais. Arquivos PNG são comumente usados para armazenar gráficos para imagens da *Web*.

Fonte: *SHARPENED PRODUCTIONS*; *BITBERRY SOFTWARE APS*, 2020, *online*, tradução nossa.

Essa ocorrência chama atenção, pois a extensão .sav foi encontrada apenas no repositório da USIL (multidisciplinar) e detém a maior porcentagem em frequência entre todas as variáveis analisadas nessa etapa. Outra observação necessária é que, assim como a UFPR, a universidade peruana (USIL) está se utilizando do *DSpace* para gerir não apenas seus conjuntos de dados, mas todos os outros documentos digitais institucionais.

Em sequência, o CIAT (disciplinar) foi o repositório observado com a maior variação entre as extensões encontradas, ultrapassando o PPBio nesse quesito. Percebe-se que os arquivos de dados científicos desse repositório são majoritariamente estruturados em formatos alfanuméricos, possuindo 60,81% de seus arquivos salvos em formatos tabelares (.csv, .xls, .xlsx, .xlt, .tab). Os repositórios do CIAT e da PUC Peru fazem parte da comunidade do *Dataverse*.

Devido ao menor peso que obtiveram dentro da metodologia utilizada, poucos conjuntos dos repositórios da PUC e MEC do Peru foram analisados, e eles são em totalidade arquivos de natureza textual e tabelar (.pdf, .csv, .xlsx, .tab). Observa-se em alguns repositórios a inexistência de padronização quanto ao uso dos programas e extensões que se

[26] Pacote Estatístico para as Ciências Sociais.

destinam ao processamento e arquivamento de dados de uma mesma natureza, como nos casos da PUC Peru e CIAT, ou nos casos do PPBio e IBICT. Por exemplo, no repositório do CIAT, os depósitos de arquivos de natureza tabelar foram feitos em 5 diversificadas extensões, 3 de programas de código fonte fechado e acesso pago e 2 de programas de código fonte aberto e acesso gratuito.

Faz-se necessário apontar que na constituição de uma política de gestão de dados científicos, deve-se buscar orientação por padrões internacionais como as 5 Estrelas para Dados Abertos, que estabelecem critérios destinados a publicações, onde “dados devem ser publicados em licença aberta e formatos não proprietários”. Assim, as instituições evitariam transtornos legais com o acesso à informação e propriedade intelectual.

Com o objetivo de distinguir as extensões encontradas em formatos proprietários e não proprietários, criaram-se os Quadros 7 e 8, respectivamente.

Quadro 7 – Extensões de formatos proprietários encontrados nos repositórios analisados.

Formato proprietário	**.doc**	**.docx**	**.ppt**	**.pptx**	**.xls**	**.xlsx**	**.xlt**	**.gdb**	**.data**	**.rar**	**.sav**
Tipo de arquivo	Documento *Microsoft Word*	Documento *Microsoft Word Open XML*	Apresentação de *PowerPoint*	Apresentação de *PowerPoint Open XML*	Planilha *Excel*	Planilha *Microsoft Excel Open XML*	*Excel Template*	Arquivo de base de dados *InterBase*	Arquivo de dados *Analysis Studio Offline*	Arquivo comprimido *WinRAR*	Arquivo de dados *SPSS*
Desenvolvedor	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Microsoft*	*Borland*	*Appricon*	Eugene Roshal	IBM
Categoria	Arquivos de texto	Arquivos de texto	Arquivos de dados	Arquivos de dados	Arquivos de planilhas	Arquivos de planilhas	Arquivos de dados	Arquivos de bancos de dados	Arquivos de dados	Arquivos compactados	Arquivos de dados
Formato	Binário	Zip	Binário	Zip	Binário	Zip	Binário	N/A	Texto	Binário	Binário

Fonte: *SHARPENED PRODUCTIONS*, 2020, *online*, tradução nossa.

Quadro 8 – Extensões de formatos não proprietários encontrados nos repositórios analisados.

Não proprietário	**.txt**	**.pdf**	**.csv**	**.tab**	**.do**	**.gpx**	**.jpg**	**.png**	**.tif**	**.zip**	**.xml**	**.rdf**
Tipo de arquivo	Arquivo de texto simples	Arquivo de formato de documento portátil	Arquivo de dados separados por abas	Arquivo de valores separados por vírgula	*Java ServLet*	Arquivo de troca por GPS	Imagem JPEG	Gráfico de rede portátil	Arquivo de imagem marcada	Arquivo compactado	Arquivo XML	Arquivo de estrutura de descrição de recurso
Desenvolvedor	N/A	*Adobe Systems*	N/A	N/A	N/A	N/A	*Joint Photographic Experts Group*	*PNG Development Group*	N/A	Phil Katz	N/A	N/A
Categoria	Arquivos de texto	Arquivos de *layout* de página	Arquivos de dados	Arquivos de texto	Arquivos *Web*	Arquivos GIS	Arquivos de imagem Raster	Arquivos de imagem Raster	Arquivos de imagem Raster	Arquivos compactados	Arquivos de dados	Arquivos de configurações
Formato	Texto	Binário	Texto	Texto	N/A	XML	Binário	Binário	Binário	Zip	XML	Texto

Fonte: *SHARPENED PRODUCTIONS*, 2020, *online*, tradução nossa.

Neles, pôde-se notar uma breve descrição das extensões e a sua devida responsabilidade intelectual por meio da aba “desenvolvedor”. Os quadros devem ser observados como exemplos e orientações a pesquisadores e gestores de repositórios no processo de manutenção de acervos digitais em sua responsabilidade.

4.2 Análise da conformidade com os Princípios FAIR

Esta subseção apresenta a segunda etapa de análise dos dados dessa investigação que se divide em outras duas fases. Nela, encontram-se os resultados da análise de conteúdo dos objetos investigados à luz dos Princípios FAIR.

4.2.1 Análise da conformidade dos repositórios com os Princípios FAIR via 5-Star Data Rating Tool

Os valores obtidos por meio do uso da ferramenta em questão estão distribuídos em duas tabelas distintas, separadas por grupos de países, como na etapa de análise anterior. Em cada uma, pode-se observar o número de amostras de cada repositório, os valores obtidos por meio das avaliações feitas, os grupos de notas divergentes e repetidas por repositório, o número de conjuntos de dados por nota, e a média aritmética ponderada das notas de todos os indivíduos das amostras por repositório.

Tabela 5 – Notas obtidas a partir da *5-Star Data Rating Tool* – repositórios brasileiros.

5-Star Data Rating Tool	**PPBio *Data Repository***	**Base de Dados Científicos da UFPR**
Amostras (n)	n = 146	n = 1
Valores obtidos	Xn={2,98; 3,11; 3,14; 3,23; 3,27; 3,29; 3,39}	Xn={2,64}
Conjunto(s) de dados científicos por nota	(p1) 5 conjuntos = {2,98} (p2) 25 conjuntos = {3,11} (p3) 14 conjuntos = {3,14} (p4) 3 conjuntos = {3,23} (p5) 88 conjuntos = {3,27} (p6) 1 conjunto = {3,29} (p7) 10 conjuntos = {3,39}	(p1) 1 conjunto = {2,64}
Média ponderada (M_p) de n	3,22	2,64

Fonte: Dados da pesquisa, 2020.

No Brasil, o repositório com maior nota também foi aquele de maior acervo, e consequentemente, maior espaço amostral entre os repositórios investigados. Como percebido na Tabela 5, o PPBio possui a maior quantidade de conjuntos próximos à maior nota, e apesar da maioria dos conjuntos receberem nota 3,27, o conjunto selecionado para

avaliação está contido no grupo p4, pois esse recebeu nota próxima à média. O repositório do PPBio também foi aquele que obteve maior variância de notas (7 grupos) entre os repositórios investigados, o que pode estar relacionado ao tempo de vida do projeto e mudanças político-técnicas; ao (des)controle da padronização no processo de depósito dos conjuntos, o qual envolve tanto depositante quanto equipe gestora do repositório; entre outros.

O Quadro 9 apresenta em texto o equivalente às respostas dadas e notas obtidas por meio das avaliações dos conjuntos selecionados em seus respectivos repositórios. As respostas dadas à ferramenta durante processo avaliativo correspondem à percepção humana quanto à realidade encontrada nos períodos de acesso aos endereços eletrônicos dos conjuntos de dados científicos.

Quadro 9 – Transcrição da avaliação dos conjuntos de dados científicos dos repositórios brasileiros na *5-Star Data Rating Tool*.

Requisitos	PPBio *Data Repository*	Base de Dados Científicos da UFPR
1) Identidade do conjunto de dados – Título* e URL	Biomassa de raízes em ecossistemas [...] https://search.dataone.org/view/liliandias.13.5	Lista de requisitos levantados […] https://bdc.c3sl.ufpr.br/handle/123456789/57 DOI: http://dx.doi.org/10.5380/bdc/38
2) Publicado – os dados estão acessíveis a usuários para além do dono ou criador?	Sim. g) Em serviço *Web* API padrão.	Sim. g) Em serviço *Web* API padrão.
3) Citável – Denotado usando um identificador formal	Sim. d) Um identificador persistente *Web* (URI).	Sim. d) Um identificador persistente *Web* (URI).
4) Descrito – marcado com metadados	Parcialmente. d) Metadados especializados (ex.: *Darwin Core*, ISO 19115/19139, perfil dados científicos schema.org).	Parcialmente. c) Metadados básicos (ex.: *Dublin Core*).
5) Encontrável* – indexado em um sistema de descoberta	Parcialmente. c) Em sistema de ampla comunidade ou de jurisdição.	Sim. d) Altamente ranqueado em índice de propósito geral (*Google*, *Bing*, etc.).
6) Carregável – representado usando um formato comum ou endossado por comunidade (ex.: padrão)	Sim. c) Múltiplos formatos padrão.	Parcialmente. b) Um formato padrão, denotado por um tipo MIME[27].
7) Usável – estruturado usando um modelo de dados ou esquema detectável e endossado por comunidade (padrão?)	Parcialmente. b) Modelo de dados ou esquema explícito, formalizado em DDL, XSD, DDI, RDFS, JSON-SCHEMA, pacote de dados ou similar.	Não. a) Nenhum esquema formal.
8) Compreensível – assistido com definições inequívocas para todos os elementos internos	Parcialmente. c) Etiquetas-padrão de comunidade (ex.: convenções CF, unidades UCUM).	Não. a) Códigos ou etiquetas de campo local.
9) Vinculado – a outros dados e definições usando identificadores públicos (ex.: URIs)	Parcialmente. b) *Links* internos a partir de um catálogo ou página de destino.	Parcialmente. b) *Links* internos a partir de um catálogo ou página de destino.
10) Licenciado – condições para reúso estão disponíveis e claramente expressas	Não. a) Nenhuma licença.	Sim. c) *Link* para uma licença padrão (ex.: *Creative Commons*).
11) Tratado – compromisso ao garantir que os dados estejam disponíveis em longo prazo	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.
12) Atualizado – parte de um programa ou série regular de coleção de dados, com disposições claras de manutenção e cronograma de atualização	Não. a) Conjunto de dados único.	Não. a) Conjunto de dados único.
13) Avaliável – acompanhado por, ou vinculado a uma avaliação da qualidade dos dados e descrição da origem e fluxo de trabalho que produziu os dados	Parcialmente. b) Declaração de linhagem em texto.	Não. a) Nenhuma informação sobre qualidade ou linhagem.
14) Confiável – acompanhado por, ou vinculado a informação sobre como os	Parcialmente. b) Estatística de uso disponível.	Não. a) Nenhuma informação sobre uso.

[27] *Multipurpose Internet Mail Extensions* (MIME) ou Extensões Multifunção para Mensagens de *Internet*.

dados foram utilizados, por quem e quantas vezes		
***Links* das avaliações**	https://oznome.csiro.au/5star?view=5e7a5b544d098386e957a18c	https://oznome.csiro.au/5star?view=5e826cc44d0983255557a210

Fonte: *5 Star Data Rating Tool*, 2020, *online*, tradução nossa. Legenda: parcialmente = não atingiu nota máxima.

Como observado em ambas as avaliações, os conjuntos não atingiram nota máxima nas questões 8, 12, 13 e 14. Quanto à questão 1, ela não possui caráter avaliativo, apenas serve como identificação dos conjuntos avaliados. A questão 12 não pôde ser respondida, pois não há informação disponível nos repositórios de que os conjuntos investigados façam parte de uma determinada série ou programa de pesquisa que sofra constante atualização.

Por ser uma questão específica, do tipo que apenas o depositante teria conhecimento para responder, ela promove uma reflexão significativa: tal informação deveria ser de preocupação também do gestor ou equipe responsável pelo processo de arquivamento do repositório, ou seja, alguém precisa se preocupar em descrevê-la durante depósito. Apesar de ser uma proposta interessante, sabe-se que as questões 12 e 14 não estão contidas explicitamente nos Princípios FAIR, tratando-se de uma adaptação dos autores por motivos desconhecidos.

Entre os conjuntos avaliados, aquele que obteve menor nota (conjunto do repositório da UFPR) foi o que recebeu maior número de respostas positivas (cinco), preenchendo completamente tais quesitos. Todavia, esse também é o conjunto que recebeu maior número de respostas negativas (cinco). Nesse processo avaliativo, as repostas “parcialmente” possuem peso, que varia entre o número de respostas possíveis, indicando que algum critério foi minimamente atendido.

Consequentemente, torna-se necessário perceber as diferenças de respostas entre os conjuntos do repositório do PPBio próximos à média e de maior nota. São elas: questão 5) opção D, o conjunto é altamente classificado em índices de uso geral, como o *Google*; questão 10) opção B, o conjunto apresenta licença descrita em texto; e questão 13) opção A, o conjunto não apresenta informação de origem/linhagem ou de qualidade. Portanto, o conjunto que obteve maior nota contou com duas respostas que atendem a todos os requisitos a mais, e uma que não atende ao requisito mínimo.

A Tabela 6 repete o mesmo processo da tabela anterior, onde se apresentam as notas referentes às avaliações de conjuntos de dados científicos dos demais repositórios sul-americanos.

Tabela 6 – Notas obtidas a partir da *5-Star Data Rating Tool* – demais repositórios sul-americanos.

5-Star Data Rating Tool	***Repositorio Institucional* USIL**	**CIAT *Dataverse***	***Portal de Datos Abiertos de la* PUC *del Perú***	***Repositorio de datos del* MEC *del Perú***
Amostras (n)	n = 59	n = 32	n = 2	n = 2
Valores obtidos	Xn={2,56; 2,66; 2,69; 2,79}	Xn={3,66; 3,76; 3,79; 3,89}	Xn ={3,59; 3,71}	Xn ={2,95; 3,05}
Conjunto(s) de dados científicos por nota	(p1) 54 conjuntos = {2,56} (p2) 1 conjunto = {2,66} (p3) 3 conjuntos = {2,69} (p4) 1 conjunto = {2,79}	(p1) 19 conjuntos = {3,66} (p2) 2 conjuntos = {3,76} (p3) 8 conjuntos = {3,79} (p4) 3 conjuntos = {3,89}	(p1) 1 conjunto = {3,59} (p2) 1 conjunto = {3,71}	(p1) 1 conjunto = {2,65} (p2) 1 conjunto = {2,75}
Média ponderada (M_p) de n	2,57	3,72	3,65	2,70

Fonte: Dados da pesquisa, 2020.

Nos casos dos repositórios colombianos e peruanos, percebe-se um certo padrão entre a variação das notas obtidas. Há 4 grupos de notas divergentes em ambos os repositórios USIL (Peru) e CIAT (Colômbia), e dois nos repositórios da PUC (Peru) e do MEC (Peru). A maior nota entre as avaliações está com o repositório CIAT *Dataverse*, seguido pelo único projeto de iniciativa privada nessa investigação, o Portal de Dados Abertos da PUC do Peru.

De um modo geral, as notas obtidas por meio da ferramenta em ênfase podem indicar que os repositórios implantados a partir do *software DSpace* possuem baixo nível FAIR. Ademais, elas podem apontar o contrário quanto aos repositórios hospedados no *Dataverse*, onde esses estão próximos de satisfazer a maioria dos critérios baseados nos Princípios FAIR estabelecidos pela ferramenta.

De forma continuada, pode-se constatar tal observação no Quadro 10, que serve como uma extensão do anterior, onde se percebe o número de questões com critérios atendidos por repositório.

Quadro 10 – Transcrição da avaliação dos conjuntos de dados científicos dos demais repositórios sul-americanos na *5-Star Data Rating Tool*.

Requisitos	***Repositorio Institucional* USIL**	**CIAT *Dataverse***	***Portal de Datos Abiertos de la* PUC *del Perú***	***Repositorio de datos del* MEC *del Perú***
1) Identidade do conjunto de dados – Título* e URL	*Competencia en tecnologías de información* [...] http://repositorio.usil.edu.pe/handle/123456789/721	*An integrated approach for understanding* [...] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QTACSN	*Estudio de Percepciones Lima* [...] http://datos.pucp.edu.pe/dataset.xhtml?persistentId=hdl:20.500.12534/3HIRBA	*Evaluación Censal de Estudiantes* […] http://datos.minedu.gob.pe/dataset/evaluaci%C3%B3n-censal-de-estudiantes-2015
2) Publicado – os dados estão acessíveis a usuários	Sim. g) Em serviço *Web* API padrão.	Sim. g) Em serviço *Web* API padrão.	Sim. g) Em serviço *Web* API padrão.	Parcialmente. d) Repositório comunitário ou institucional.

para além do dono ou criador?				
3) Citável – Denotado usando um identificador formal	Sim. d) Um identificador persistente *Web* (URI).	Sim. d) Um identificador persistente *Web* (URI).	Sim. d) Um identificador persistente *Web* (URI).	Parcialmente. c) Endereço *Web* (URL – não garantido estável).
4) Descrito – marcado com metadados	Parcialmente. c) Metadados básicos (ex.: *Dublin Core*).	Parcialmente. c) Metadados básicos (ex.: *Dublin Core*).	Parcialmente. c) Metadados básicos (ex.: *Dublin Core*).	Sim. e) Metadados ricos usando múltiplos vocabulários padrões RDF (ex.: DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)
5) Encontrável – indexado em um sistema de descoberta	Sim. d) Altamente ranqueado em índice de propósito geral (*Google*, *Bing*, etc.).	Sim. d) Altamente ranqueado em índice de propósito geral (*Google*, *Bing*, etc.).	Sim. d) Altamente ranqueado em índice de propósito geral (*Google*, *Bing*, etc.).	Sim. d) Altamente ranqueado em índice de propósito geral (*Google*, *Bing*, etc.).
6) Carregável – representado usando um formato comum ou endossado por comunidade (ex.: padrão)	Não. a) Formato customizado.	Parcialmente. b) Um formato padrão, denotado por um tipo MIME.	Parcialmente. b) Um formato padrão, denotado por um tipo MIME.	Sim. c) Múltiplos formatos padrões.
7) Usável – estruturado usando um modelo de dados ou esquema detectável e endossado por comunidade (padrão?)	Não. a) Nenhum esquema formal.	Sim. c) Modelo de dados ou esquema compartilhado por comunidade, disponível a partir de uma localização padrão.	Sim. c) Modelo de dados ou esquema compartilhado por comunidade, disponível a partir de uma localização padrão.	Parcialmente. b) Modelo de dados ou esquema explícito, formalizado em DDL, XSD, DDI, RDFS, JSON-SCHEMA, pacote de dados ou similar.
8) Compreensível – assistido com definições inequívocas para todos os elementos internos	Não. a) Códigos ou etiquetas de campo local.	Parcialmente. d) Alguns campos com *links* para definições geridas externamente.	Parcialmente. c) Etiquetas-padrão de comunidade (ex.: convenções CF, unidades UCUM).	Não. a) Códigos ou etiquetas de campo local.
9) Vinculado – a outros dados e definições usando identificadores públicos (ex.: URIs)	Parcialmente. b) *Links* internos a partir de um catálogo ou página de destino.	Sim. c) *Links* externos a definições e dados relacionados.	Parcialmente. b) *Links* internos a partir de um catálogo ou página de destino.	Parcialmente. b) *Links* internos a partir de um catálogo ou página de destino.
10) Licenciado – condições para reúso estão disponíveis e claramente expressas	Não. a) Nenhuma licença.	Sim. c) *Link* para uma licença padrão (ex.: *Creative Commons*).	Sim. c) *Link* para uma licença padrão (ex.: *Creative Commons*).	Parcialmente. b) Licença descrita em texto.
11) Tratado – compromisso ao garantir que os dados estejam disponíveis em longo prazo	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.	Parcialmente. c) Repositório público ou institucional (ex.: CKAN, *GitHub*).
12) Atualizado – parte de um programa ou série regular de coleção de dados, com disposições claras de manutenção e cronograma de atualização	Não. a) Conjunto de dados único.	Não. a) Conjunto de dados único.	Não. a) Conjunto de dados único.	Não. a) Conjunto de dados único.
13) Avaliável – acompanhado por, ou vinculado a uma avaliação da qualidade dos dados e descrição da origem e fluxo de trabalho que produziu os dados	Não. a) Nenhuma informação sobre qualidade ou linhagem.	Parcialmente. b) Declaração de linhagem em texto.	Parcialmente. b) Declaração de linhagem em texto.	Não. a) Nenhuma informação sobre qualidade ou linhagem.
14) Confiável – acompanhado por, ou vinculado a informação sobre como os dados foram utilizados, por quem e quantas vezes	Parcialmente. b) Estatística de uso disponível.	Parcialmente. b) Estatística de uso disponível.	Parcialmente. b) Estatística de uso disponível.	Não. a) Nenhuma informação sobre uso.

***Links* das avaliações**	https://oznome.csiro.au/5star?view=5e7fb79e4d0983e9bf57a1fd	https://oznome.csiro.au/5star?view=5e8249134d0983293c57a205	https://oznome.csiro.au/5star?view=5e8253564d09836d0257a20c	https://oznome.csiro.au/5star?view=5e8263d74d0983475957a20e

Fonte: Dados da pesquisa, 2020.

Frisa-se que as respostas dispostas nos quadros podem ser diferentes dependendo da escolha do conjunto referência, isso porque se constatou a existência de divergências entre os conjuntos arquivados em um mesmo repositório. Como explicado, os conjuntos tidos como referência nessa fase são aqueles próximos à média das notas.

Com o reconhecimento de que somente 11 das 14 questões se referem aos Princípios FAIR, atenta-se ao conjunto de dados científicos do repositório CIAT, pois ele atende integralmente aos critérios de 7 delas (63%). No entanto, como observado anteriormente, a nota se eleva com critérios parcialmente atendidos, apesar de não se saber como os pontos são distribuídos. Essa informação assiste os conjuntos bem avaliados a se aproximarem do topo, ou seja, indica que seu nível FAIR está mais alto que o esperado (tendo como base a avaliação feita na *5-Star Data Rating Tool*). Em exercício investigativo, descobriu-se que algumas questões (entre elas, a 12 e a 14) possuem maior peso na avaliação dos conjuntos, pois elas contribuem com 0,50 de nota à totalidade dos pontos (estrelas).

Para que cem por cento dos critérios das questões 8, 12, 13 e 14 fossem atendidos, os conjuntos precisariam ter ou ser, respectivamente: todos os campos vinculados a definições padrão geridas externamente; parte de série de dados com atualizações regulares; rastreamento formal de proveniência (por exemplo, PROV-O); e claramente endossado por estrutura ou organização conceituada.

A próxima subseção detalha os motivos e justificativas das respostas obtidas nessa fase em dois quadros criados a partir da análise qualitativa de conteúdo dos repositórios e seus conjuntos de dados científicos investigados. Todavia, como já percebido, não há relação de equivalência entre a fase de avaliação semiautomática por computador e humana, pois os meios divergem, mesmo que minimamente.

4.2.2 Análise qualitativa da conformidade dos repositórios com os Princípios FAIR

No intuito de avaliar conjuntos de dados científicos e consequentemente seus repositórios, buscou-se nesse trabalho um meio de selecionar amostras para que se justificasse metodologicamente o porquê da análise de um conjunto em detrimento doutro. Assim, chegou-se à resposta de como separar para avaliar, descrita no Capítulo 3.

Em posse dos resultados da etapa anterior, especificamente, dos valores obtidos pelas avaliações das amostras em ferramenta semiautomática, decidiu-se então por analisar

qualitativamente os conjuntos de modo geral, não se baseando apenas em um conjunto referência. Dessa forma, a análise foi feita à luz da leitura dos Princípios FAIR, onde cada detalhe das proposições foi conferido e comparado à realidade desses repositórios com o proposto.

A partir daí, obteve-se o resultado apresentado em dois quadros que contêm praticamente todo o conteúdo de interesse dessa subseção. Os quadros são, portanto, um resumo da avaliação feita. Outros detalhes percebidos durante análise dos objetos de estudo complementam as informações dos quadros.

O Quadro 11 apresenta os resultados das análises qualitativas dos conjuntos e repositórios brasileiros, seguindo o padrão comparativo estabelecido na dissertação. Nessa fase de avaliação, os critérios atendidos foram calculados, resultando uma nota entre 0 e 15, onde zero significa nenhum critério atendido, e quinze, todos os critérios atendidos. Não há pontuação aos critérios atendidos parcialmente, dessa maneira, o resultado será sempre inteiro. Obteve-se então uma escala percentual simples, na qual 15 critérios são divididos por 100%, deixando cada critério atendido equivalente a 6,6%. Essa escala se conecta com a ideia do nível FAIR dos dados e metadados, tão citado na bibliografia da área.

Quadro 11 – Conformidade dos repositórios brasileiros com os Princípios FAIR.

Princípios FAIR		**PPBio *Data Repository (DataONE)***	**Base de Dados Científicos da UFPR (*DSpace*)**
Findable **(Encontráveis)**	F1	Atende totalmente: aos conjuntos de dados e metadados são atribuídos identificadores, contudo nem todos os registros recebem identificadores persistentes únicos como DOI, *Handle*, URN, etc. Ex.: urn:uuid:602d75d3-5348-4950-b9ae-fcb9f0f64b66; PPBioAmOc.534.4; liliandias.38.2; drucker.3.9; menger.35.3	Atende totalmente: aos conjuntos de dados e metadados são atribuídos identificadores persistentes únicos. Nesse caso, em um mesmo registro, metadados descrevem número de DOI, enquanto seu URI recebe número de *Handle*. Ex.: http://dx.doi.org/10.5380/bdc/38 https://bdc.c3sl.ufpr.br/handle/123456789/57
	F2	Atende totalmente: a descrição de grande parte dos conjuntos é exaustiva no que tange à contextualização dos dados (*rich metadata*).	Não atende: a descrição de grande parte dos conjuntos é insuficiente no que tange à contextualização dos dados (*poor metadata*).
	F3	Atende totalmente: os metadados apresentam explicitamente os identificadores persistentes dos dados em seu registro, anotados na *tag packageId* em arquivo XML.	Atende totalmente: os metadados apresentam explicitamente os identificadores persistentes dos dados em seu registro, anotados na *tag DC.identifier* em código fonte.
	F4	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.
Accessible **(Acessíveis)**	A1	Não atende: o identificador não disponibiliza acesso aos (meta)dados por *link*, pois não possui protocolo de comunicação padronizado (http, https, ftp, etc.).	Atende totalmente: os conjuntos e seus (meta)dados podem ser recuperados pelo seu DOI por acesso a *links* e a comunicação se dá por protocolo de tranferência de hipertexto - *Hypertext Transfer Protocol* (HTTP). Ex.: http://dx.doi.org/10.5380/bdc/38.
	A1.1	Não atende: não há protocolo de comunicação padronizado presente na configuração dos identificadores persistentes atribuídos aos conjuntos de dados do repositório.	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente implementável.
	A1.2	Não atende: não há protocolo de autenticação e autorização explícito. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.	Não atende: o HTTP não permite procedimento de autenticação e autorização por meio de conexão criptografada. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.

	A2	Atende totalmente: aproximadamente ¼ de todos os registros encontrados apenas disponibiliza acesso aos metadados. Dessa forma, pode-se dizer que os metadados estão disponíves mesmo quando os dados não estão. Obs.: para acesso a dados embargados ou não publicados, não há protocolo de autorização formalizado explicitamente.	Atende totalmente: o *DSpace* possibilita o embargo temporário aos dados via solicitação do autor/criador, no entanto mantém o acesso a seus metadados. Obs.: para acesso a dados embargados ou não publicados, não há protocolo de autorização formalizado explicitamente.
***Interoperable* (Interoperáveis)**	I1	Atende totalmente: o modelo dos dados é composto pela EML. Todos os conjuntos de dados do repositório possuem arquivos XML que contêm anotações em EML (servem à exportação dos metadados).	Atende totalmente: o modelo dos dados é composto pelo esquema de metadados *Dublin Core,* padrão no *DSpace*. Obs.: não há opção de exportação de metadados.
	I2	Não atende: não há uso de vocabulário controlado na descrição dos dados.	Não atende: não há uso de vocabulário controlado na descrição dos dados.
	I3	Atende totalmente: os (meta)dados possuem referências qualificadas a outros (meta)dados por meio de texto ou *links* a trabalhos relacionados.	Não atende: os (meta)dados não possuem referências qualificadas a outros (meta)dados. Não há menção ou *links* a trabalhos relacionados.
***Reusable* (Reusáveis)**	R1	Atende totalmente: os dados são descritos de forma exaustiva e específica em relação a seu conteúdo e contexto em que foram gerados. Além do título, identificador, resumo/descrição e palavras-chave, pode-se encontrar descrição da região geográfica, da cobertura temporal, do alcance taxonômico, dos métodos e amostragem, etc.	Não atende: há pouca descrição quanto ao conteúdo e contexto dos dados. Basicamente, há pouco mais que título, autor/criador, identificador, resumo e palavras-chave.
	R1.1	Atende totalmente: há a descrição em texto da licença *Creative Commons*.	Atende totalmente: os conjuntos de dados científicos são acompanhados de arquivo em RDF onde se tem acesso a anotações relativas à licença *Creative Commons*. Há também descrição em texto e *link* de acesso à licença.
	R1.2	Atende totalmente: além dos nomes dos autores/criadores dos dados, pode-se encontrar seus contatos, sua instituição de origem, equipe do projeto, órgão financiador, entre outros.	Não atende: não há detalhamento dos responsáveis e/ou envolvidos na origem dos dados.
	R1.3	Atende totalmente: tratando-se de um repositório de domínio (Ecologia), entende-se que a EML e o nível descritivo encontrado atendem ao padrão de comunidade de domínio. O repositório possui uma seção de Melhores Práticas, que orienta a gestão dos dados a sua padronização.	Atende parcialmente: tratando-se de um repositório de comunidade acadêmica, multidisciplinar por natureza, entende-se que o *Dublin Core* e o nível descritivo encontrado atendem parcialmente ao padrão de comunidade de domínio. O repositório não possui uma seção de Melhores Práticas.
Critérios atendidos		11 (72,6%)	8 (52,8%)

Fonte: Dados da pesquisa, 2020.

De forma objetiva, o quadro apresenta os motivos que fazem cada critério ser atendido ou não, assim como o número de critérios atendidos, sustentando seu propósito comparativo. Como observado, dentro da escala percentual, os repositórios do PPBio e da UFPR obtiveram nível FAIR de 72,6% e 52,8%, respectivamente. Pelos resultados encontrados em âmbito nacional, entende-se que o repositório do PPBio tem maior maturidade frente à gestão de dados científicos. Um dos motivos pode estar relacionado à natureza desse repositório, pois se trata de um projeto disciplinar atrelado à iniciativa *DataONE* e a um programa de pós-graduação que se destina apenas aos dados (não-híbrido).

Na primeira fase de avaliação dos conjuntos do PPBio não foi possível acessá-los por meio de seu endereço eletrônico, sendo preciso se conectar à plataforma do *DataONE* e buscar o conjunto por seu identificador ou outro descritor. Pelo *Google*, podia-se recuperar conjuntos via título, autor(es) e *abstract*, menos por identificador. Em alguns casos, recuperava-se apenas por um ou outro metadado. Como visto em tabela, os identificadores

não possuíam protocolo de comunicação padronizado, o que incapacitou o acesso aos conjuntos via *links*.

Durante o período de reavaliação, verificou-se que houve atualização no repositório, posto que se tornou possível recuperar conjuntos no *Google* por seus identificadores. Além disso, observou-se que seus URIs também foram atualizados, o que permitiu acesso direto às páginas dos conjuntos. Por conseguinte, percebeu-se que o repositório foi alimentado entre as duas etapas descritas, recebendo mais depósitos.

Apesar de atender a pouco mais da metade dos requisitos, o repositório da UFPR não teve problemas relacionados a acesso. Também não se observou inconsistências ou manutenções no sistema, tampouco se percebeu atualização de qualquer natureza (ex.: novos depósitos). Certamente há algum tipo de investimento sendo feito ao projeto, pois em âmbito nacional está à frente das demais universidades brasileiras. Porém, os meios utilizados e a divisão entre tarefas complexas e distintas num ambiente compartilhado dificultam o desenvolvimento de um serviço especializado orientado a dados científicos.

O Quadro 12 expõe a última parcela dos resultados, a qual diz respeito aos demais repositórios sul-americanos. Fazem parte desse grupo: 3 instituições de ensino, e uma instituição de administração pública.

Quadro 12 – Conformidade dos demais repositórios sul-americanos com os Princípios FAIR.

Princípios FAIR		***Repositorio Institucional* USIL (*DSpace*)**	**CIAT (*Dataverse*)**	***Portal de Datos Abiertos de la PUC del Perú* (*Dataverse*)**	***Repositorio de datos del* MEC *del Perú* (DKAN)**
***Findable* (Encontráveis)**	F1	Atende totalmente: aos conjuntos de dados e metadados são atribuídos identificadores persistentes únicos. Nesse caso, *Handle*. Ex.: http://repositorio.usil.edu.pe/handle/USIL/9511	Atende totalmente: aos conjuntos de dados e metadados são atribuídos identificadores persistentes únicos. Nesse caso, DOI. Ex.: doi:10.7910/DVN/PIOKQZ	Atende totalmente: : aos conjuntos de dados e metadados são atribuídos identificadores persistentes únicos. Nesse caso, *Handle*. Ex.: hdl:20.500.12534/3HIRBA	Atende parcialmente: aos conjuntos de dados e metadados são atribuídos identificadores, contudo não são persistentes únicos como DOI, *Handle*, URN, etc. Ex.: 3f353785-035b-455b-917c-4be49623b025
	F2	Não atende: a descrição de grande parte dos conjuntos é insuficiente no que tange à contextualização dos dados (*poor metadata*).	Atende totalmente: a descrição de grande parte dos conjuntos é exaustiva no que tange à contextualização dos dados (*rich metadata*).	Atende totalmente: a descrição de grande parte dos conjuntos é exaustiva no que tange à contextualização dos dados (*rich metadata*).	Não atende: a descrição de grande parte dos conjuntos é insuficiente no que tange à contextualização dos dados (*poor metadata*).
	F3	Atende totalmente: os metadados apresentam explicitamente os identificadores persistentes dos dados em seu registro, anotados na *tag DC.identifier* em código fonte.	Atende totalmente: os metadados apresentam explicitamente os identificadores persistentes dos dados em seu registro, anotados na *tag DC.identifier* em código fonte.	Atende totalmente: os metadados apresentam explicitamente os identificadores persistentes dos dados em seu registro, anotados na *tag DC.identifier* em código fonte.	Atende totalmente: os metadados apresentam explicitamente os identificadores dos dados em seu registro, anotados na *tag dcterms:identifier* em código fonte.
	F4	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.	Atende totalmente: os metadados estão registrados e indexados de forma a possibilitar a devida recuperação dos conjuntos de dados científicos via *Google*. Obs.: Repositório indexado na base RE3DATA.
***Accessible* (Acessíveis)**	A1	Atende totalmente: os conjuntos e seus	Atende totalmente: os conjuntos e seus	Atende totalmente: os conjuntos e seus	Não atende: o identificador não disponibiliza acesso aos

		(meta)dados podem ser recuperados por acesso a *links* e a comunicação se dá por protocolo de tranferência de hipertexto - *Hypertext Transfer Protocol* (HTTP). Ex.: http://repositorio.usil.edu.pe/handle/USIL/9511.	(meta)dados podem ser recuperados por acesso a *links* de DOI e a comunicação se dá por protocolo de tranferência de hipertexto seguro - *Hypertext Transfer Protocol Secure* (HTTPS), extensão do HTTP. Ex.: https://doi.org/10.7910/DVN/PIOKQZ.	(meta)dados podem ser recuperados por acesso a *links* de *Handle* e a comunicação se dá por protocolo de tranferência de hipertexto seguro - *Hypertext Transfer Protocol Secure* (HTTPS), extensão do HTTP. Ex.: https://hdl.handle.net/20.500.12534/3HIRBA.	(meta)dados por *link*, pois não possui protocolo de comunicação padronizado (http, https, ftp, etc.).
	A1.1	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente implementável.	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente implementável.	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente implementável.	Não atende: não há protocolo de comunicação padronizado presente na configuração dos identificadores persistentes atribuídos aos conjuntos de dados do repositório.
	A1.2	Não atende: o HTTP não permite procedimento de autenticação e autorização por meio de conexão criptografada. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.	Atende totalmente: o HTTPS permite procedimento de autenticação e autorização por meio de uma conexão criptografada que verifica a autenticidade do servidor e do cliente via certificados digitais. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.	Atende totalmente: o HTTPS permite procedimento de autenticação e autorização por meio de uma conexão criptografada que verifica a autenticidade do servidor e do cliente via certificados digitais. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.	Não atende: não há protocolo de autenticação e autorização explícito. Obs.: o sistema não solicita a criação de conta de usuário para dar acesso aos (meta)dados, mas é possível fazê-lo.
	A2	Atende totalmente: o *DSpace* possibilita o embargo temporário aos dados via solicitação do autor/criador, no entanto mantém o acesso a seus metadados. Obs.: para acesso a dados embargados ou não publicados, não há protocolo de autorização formalizado explicitamente.	Atende totalmente: o *Dataverse* possibilita o embargo temporário aos dados via solicitação do autor/criador, no entanto mantém o acesso a seus metadados. Obs.: para acesso a dados embargados ou não publicados, não há protocolo de autorização formalizado explicitamente.	Atende totalmente: o *Dataverse* possibilita o embargo temporário aos dados via solicitação do autor/criador, no entanto mantém o acesso a seus metadados. Obs.: para acesso a dados embargados ou não publicados, não há protocolo de autorização formalizado explicitamente.	Não atende: o DKAN possibilita a criação de conjuntos de dados sem a necessidade da sua publicação. Assim, eles podem ficar arquivados no sistema por tempo indeterminado, sem acesso. Entretanto, não há opção de acesso aos metadados com embargo temporário aos dados via solicitação do autor/criador.
Interoperable **(Interoperáveis)**	I1	Atende totalmente: o modelo dos dados é composto pelo esquema de metadados *Dublin Core,* padrão no *DSpace*. Obs.: não há opção de exportação de metadados.	Atende totalmente: o modelo dos dados é composto pelo esquema de metadados *Dublin Core,* padrão no *Dataverse*. Obs.: é possível exportar os metadados em *Dublin Core*, DDI, DDI HTML *Codebook*, *DataCite*, JSON, OAI_ORE, *Open*AIRE, e Schema.org JSON-LD.	Atende totalmente: o modelo dos dados é composto pelo esquema de metadados *Dublin Core,* padrão no *Dataverse*. Obs.: é possível exportar os metadados em *Dublin Core*, DDI, *DataCite*, JSON, OAI_ORE, e Schema.org JSON-LD.	Atende totalmente: o modelo dos dados é composto pelos esquemas *Dublin Core,* FOAF, SIOC, SKOS, OWL, DCAT. Obs.: não há opção de exportação de metadados.
	I2	Não atende: não há uso de vocabulário controlado na descrição dos dados.	Atende parcialmente: o repositório utiliza o tesauro AGROVOC (*SKOS schema*), que cobre as áreas de interesse da *Food and Agriculture Oganization* (FAO) das Nações Unidas (AIMS, 2020, *online*, tradução nossa). Ele alimenta as palavras-chave e as classificações de tópicos dos conjuntos depositados no repositório. Todavia, o vocabulário controlado AGROVOC não possui identificador persistente. Obs.: é possível acessar as tabelas dos termos autorizados por *links* disponíveis nos registros.	Não atende: não há uso de vocabulário controlado na descrição dos dados.	Não atende: não há uso de vocabulário controlado na descrição dos dados.

	I3	Não atende: os (meta)dados não possuem referências qualificadas a outros (meta)dados. Não há menção ou *links* a trabalhos relacionados.	Atende totalmente: os (meta)dados possuem referências qualificadas a outros (meta)dados por meio de texto ou *links* a trabalhos relacionados.	Atende totalmente: os (meta)dados possuem referências qualificadas a outros (meta)dados por meio de texto ou *links* a trabalhos relacionados.	Não atende: os (meta)dados não possuem referências qualificadas a outros (meta)dados. Não há menção ou *links* a trabalhos relacionados.
Reusable **(Reusáveis)**	R1	Não atende: há pouca descrição quanto ao conteúdo e contexto dos dados. Basicamente, há pouco mais que título, autor/criador, identificador, resumo e palavras-chave.	Atende totalmente: os dados são descritos de forma exaustiva e específica em relação a seu conteúdo e contexto em que foram gerados. Além do título, identificador, resumo/descrição e palavras-chave, pode-se encontrar descrição da cobertura geográfica e temporal, do tipo dos dados, publicações e conjuntos de dados relacionados, universo da pesquisa, entre outros.	Atende totalmente: os dados são descritos de forma exaustiva e específica em relação a seu conteúdo e contexto em que foram gerados. Além do título, identificador, resumo/descrição e palavras-chave, pode-se encontrar descrição do universo da pesquisa, do coletor dos dados, do procedimento amostral, do tamanho da amostra alvo, da estimativa de erro da amostra, entre outros.	Não atende: há pouca descrição quanto ao conteúdo e contexto dos dados. Basicamente, há pouco mais que título, autor/criador e identificador.
	R1.1	Atende totalmente: há menção à licença *Open Database License* em texto e é possível acessá-la por *link*.	Atende totalmente: na aba *Terms*, é possível acessar o texto e *links* para as normas da comunidade e a licença *Creative Commons*.	Atende totalmente: na aba *Terms*, é possível acessar o texto da *CCO Public Domain Dedication* e *link* para as normas da comunidade.	Atende totalmente: há menção à licença *Open Data Commons Attribution License*.
	R1.2	Não atende: não há detalhamento dos responsáveis e/ou envolvidos na origem dos dados.	Atende totalmente: além dos nomes dos autores/criadores dos dados, pode-se encontrar seus números de ORCID, sua instituição de origem, produtor, distribuidor, depositador dos dados, informação de financiamento, datas, etc.	Atende totalmente: além dos nomes dos autores/criadores dos dados, pode-se encontrar sua instituição de origem, produtor, contribuidor, depositador dos dados, informação de financiamento, datas, etc.	Não atende: não há detalhamento dos responsáveis e/ou envolvidos na origem dos dados.
	R1.3	Atende parcialmente: tratando-se de um repositório de comunidade acadêmica, multidisciplinar por natureza, entende-se que o *Dublin Core* e o nível descritivo encontrado atendem parcialmente a padrões de comunidades de domínio. O repositório não possui uma seção de Melhores Práticas.	Atende totalmente: tratando-se de um repositório de domínio (Agricultura), entende-se que o uso do vocabulário controlado AGROVOC e o nível descritivo encontrado atendem ao padrão de comunidade de domínio. O repositório possui uma seção de Melhores Práticas, que orienta a gestão dos dados a sua padronização.	Atende totalmente: tratando-se de um repositório de domínio (Ciências Sociais), entende-se que o nível descritivo encontrado atende ao padrão de comunidade de domínio. O repositório possui uma seção de Melhores Práticas, que orienta a gestão dos dados a sua padronização.	Não atende: tratando-se de um repositório de dados governamentais, entende-se que pode não haver uma comunidade ativa envolvida no projeto. O nível descritivo encontrado é baixo e padronizado, como se depositado por quem não tivesse informação suficiente para fazê-lo, mas que segue um protocolo. O repositório não possui uma seção de Melhores Práticas.
Critérios atendidos		8 (52,8%)	14 (92,4%)	14 (92,4%)	4 (26,4%)

Fonte: Dados da pesquisa, 2020.

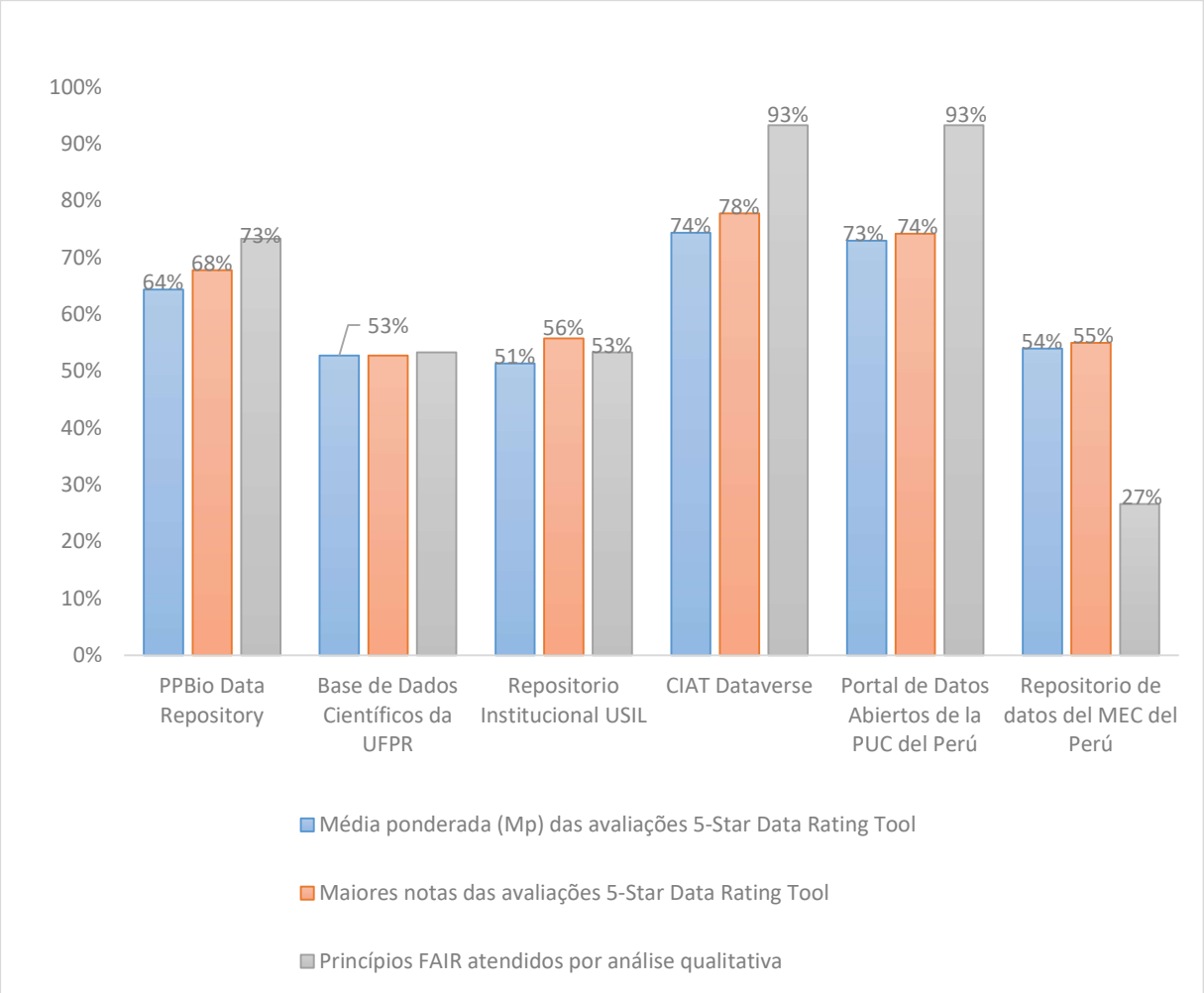
A partir da análise dos dados apresentados nos quadros, constatou-se “empate técnico” entre os repositórios da UFPR e do USIL, isso porque ambos atenderam totalmente a 8 critérios e a 1 parcialmente. Como se sabe, os dois são operacionalizados pelo *DSpace*, o que pode justificar os resultados obtidos. Deve-se lembrar que na fase da avaliação semiautomática com o *5-Star Data Rating Tool*, os repositórios implantados pelo *software DSpace* obtiveram as menores notas entre os avaliados. Ademais, os dois atingiram entre 51 e 56 por cento de nível FAIR nas duas etapas de avaliação.

Outra demonstração de que não há coincidência na escolha de *software* está nos resultados alcançados pelos repositórios do CIAT e da PUC Peru, hospedados pelo *Dataverse*. O projeto da Universidade de *Harvard* nasce tendo como principal objetivo a

gestão e curadoria de dados científicos. Atendendo totalmente a 14 critérios, ambos obtiveram resultados satisfatórios, especialmente o repositório do CIAT, que ficou próximo de cumprir todos os requisitos (critério I2 quase atingido).

O repositório do MEC Peru recebeu a menor nota nessa fase de avaliação, se distanciando da anterior pela metade, como mostra o Gráfico 1.

Gráfico 1 – Nível FAIR dos repositórios sul-americanos.



Fonte: Dados da pesquisa, 2020.

Trata-se de um repositório de dados governamentais abertos, nos moldes do portal da transparência do governo federal brasileiro, que serve como fonte de prestação de contas do governo peruano à população. O repositório de dados abertos do MEC Peru foi implantado pelo *software* DKAN, “plataforma gratuita de código e dados abertos que dá liberdade a organizações e indivíduos para publicar e consumir informação estruturada”. (DKAN, 2020, *online*, tradução nossa).

O DKAN não é apenas semelhante ao CKAN em seu nome – sua comunidade de usuários/clientes é majoritariamente (senão totalmente) composta por organizações governamentais, isso de acordo com o informado em seu *Website*.

4.3 Discussão dos resultados

A começar pelo primeiro objetivo estabelecido para esta investigação, partiu-se do ponto comum entre pesquisas de caráter exploratório, onde se obteve o primeiro contato com o universo a ser investigado. Os pressupostos partiam da ideia de que as abordagens aos assuntos pesquisados poderiam ser incipientes em nível continental. De certa forma, os achados corroboram tal impressão, uma vez que o número de projetos de repositórios de dados científicos se demonstrou baixo quando comparado à quantidade de instituições de pesquisa existentes em toda a América do Sul. No entanto, sob uma ótica otimista, pode-se dizer que os três países em questão (Brasil, Colômbia e Peru) estão à frente dos demais vizinhos nesse quesito.

Por meio da metodologia aplicada à seleção dos objetos de estudo, foi possível observar que Brasil e Peru estão em situações parecidas, contando com dois repositórios de dados científicos ativos, diferenciando-se entre os produtos de *software* utilizados em sua implementação e o número de conjuntos depositados, pois como visto, o repositório do PPBio tem maior volume, o que pode estar diretamente relacionado a sua maturidade (tempo, investimentos, envolvimento e consciência da comunidade, política de gestão, entre outros).

Destaca-se também que o PPBio está no grupo de repositórios de domínio, assim como o do CIAT e o da PUC Peru, contudo hospedado e gerenciado pelo *DataONE*. A partir das avaliações e análises feitas, percebeu-se que os repositórios implantados por meio do *software Dataverse* estão em maior conformidade com os Princípios FAIR. Todavia, o *Dataverse* tem caráter multidisciplinar, enquanto que o *DataONE* serve apenas à comunidade que se destina. Em termos de qualidade de serviço, esse pode ser um ponto em que o *DataONE* se difere positivamente, pois une toda uma comunidade (Iniciativa *DataONE*) em prol de um objetivo em comum, fortalecendo assim sua identidade.

Em relação aos conjuntos de dados científicos investigados, notou-se que sua natureza se concentra em dados textuais e numéricos, salvos em arquivos de texto e em tabelas, respectivamente. Como observado por Sales e Sayão (2019, p. 41), tabelas têm natureza visual, entretanto, seus dados são resultados de levantamentos, fórmulas, equações, entre outros, ou seja, de natureza numérica.

Por meio das análises dos formatos e extensões dos arquivos de dados, percebeu-se que os conjuntos de dados podem ser tanto homogêneos (um ou mais arquivos salvos em um único formato e extensão, ex.: formato de imagem em .jpg) ou heterogêneos (arquivos salvos em diferentes formatos e extensões, ex.: mesmo formato de imagem salvo em .jpg e .tiff) em sua composição. Apurou-se também que algumas extensões possibilitam tanto a identificação da natureza quanto do domínio e do conteúdo dos dados, como observado nas extensões .gpx e .gdb, que se referem a dados de localização geográfica, logo,

de natureza alfanumérica. Há crescente necessidade de se descrever a natureza dos dados, assim como os formatos ou extensões de seus arquivos. Esse tipo de metadado descritivo seria valioso a potenciais usuários, pois permitiria obter maior compreensão do contexto dos dados com foco em seu reúso.

Os achados corroboram as afirmações de Borgman, Scharnhorst e Golshan (2019, p. 889) em que arquivos de dados digitais não são entidades monolíticas. Alguns coletam apenas dados de certos tipos e formatos, como sequências de genômas para a pesquisa biológica ou dados de *surveys* para as ciências econômicas e sociais. Outros são mais genéricos, coletando documentos textuais, imagens estáticas ou em movimento, áudio e outros tipos de dados (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 889, tradução nossa).

Não foi objetivo dessa investigação avaliar a *5-Star Data Rating Tool* e os Princípios FAIR, mas utilizá-los como meio a um fim. Pelo processo avaliativo, pôde-se perceber que a ferramenta semiautomática de autoavaliação não se baseia integralmente nesses princípios, sofrendo adaptação pelos autores por motivos desconhecidos. Apesar disso, em 3 dos 5 repositórios que buscam gerir e curar dados científicos (obs.: o repositório do MEC Peru não se destina a tal), os resultados indicam alinhamento entre a ferramenta e os princípios estudados.

Portanto, entende-se que a referida ferramenta não substitui a leitura dos princípios para melhor entendimento de seus critérios e proposições, ou seja, não substitui a avaliação humana. Em situação de composição de um projeto de repositório de dados científicos, sua leitura se torna essencial e indispensável à equipe envolvida. Isso não quer dizer que a ferramenta seja de pouca serventia ao processo de adaptação FAIR. Na verdade, ela servirá como orientação aos detentores dos dados (pesquisadores depositantes), uma vez que a avaliação proposta não exige conhecimento prévio sobre os Princípios FAIR, assim como permite seu usuário verificar a situação em que seus dados se encontram, caso seu objetivo esteja no arquivamento e no compartilhamento desses ativos.

5 CONSIDERAÇÕES FINAIS

Primeiramente, acusa-se o cumprimento de todos os objetivos traçados. À princípio, havia uma enorme quantidade de dúvidas e incertezas sobre os assuntos abordados e possíveis resultados. Apesar disso, a proposta de pesquisa, que inicialmente pareceu nebulosa, mostrou-se promissora.

Os repositórios encontrados e aptos à investigação foram o PPBio *Data Repository* (Brasil); a Base de Dados Científicos da Universidade Federal do Paraná (Brasil); o CIAT *Dataverse* (Colômbia); o *Portal de Datos Abiertos de la Pontificia Universidad Católica del Perú* (Peru); o *Repositorio de datos del Ministerio de Educación del Perú* (Peru); e o *Repositorio Institucional* USIL (Peru). Os demais repositórios eliminados passaram por algum problema antes da fase de análise, o que justifica a lista supracitada. O repositório do MEC Peru poderia ter sido eliminado antecipadamente, porém sua inclusão se tornou relevante aos resultados obtidos.

Os programas por trás dos repositórios investigados que servem à gestão e curadoria de dados científicos são o *Morpho* (*DataONE*), o *DSpace*, e o *Dataverse*. Como percebido e declarado na literatura, cada um serve a um propósito e a uma ou mais comunidades. O *Morpho* se destina à comunidade compartilhada entre as Ciências Biológicas e Ecologia, foco em dados científicos; o *DSpace* (multidisciplinar), à preservação digital institucional, foco em documentos bibliográficos; o *Dataverse* (multidisciplinar), a comunidades compartilhadas entre diversos campos de pesquisa, foco em dados científicos.

Os repositórios em maior conformidade com os Princípios FAIR foram aqueles estabelecidos mediante o uso do *software Dataverse*. Para além de tal constatação, pressupõe-se que suas notas foram similarmente reflexo do fator humano, ou seja, dos serviços desempenhados por profissionais responsáveis.

De forma a atender a padrões científicos, apresentam-se as limitações e contribuições dessa investigação, assim como perspectivas para futuras pesquisas, partindo-se do ponto do encerramento desta.

A maior parte das limitações existentes no desenvolvimento dessa investigação se encontra no processo metodológico utilizado. A declaração se justifica pelos meios de execução do levantamento dos repositórios de dados científicos, uma vez que se desconhecia caminhos alternativos para fazê-lo. Não se pode afirmar que os repositórios encontrados são os únicos existentes, entretanto, a partir das buscas, pode-se dizer que, caso existam, estão submersos em um vasto oceano de dados e informações na *Web*, e portanto, precisam indexar com fim em sua recuperação.

Outra limitação anteriormente apontada está no universo investigado, onde se encontram mais de mil conjuntos de dados científicos. Em uma situação ideal, tendo devido acesso a recursos suficientes, todos os conjuntos seriam analisados em sua íntegra. Isso proporcionaria uma avaliação mais precisa da realidade desses e, consequentemente, de seus repositórios hospedeiros.

Apesar dessa investigação se caracterizar como uma pesquisa qualitativa indutiva, não há como generalizar os resultados obtidos (vide problema da indução), posto que o universo investigado se refere a um pequeno recorte do todo, ou seja, da regionalização dos objetos estudados. Essa situação abre oportunidades para futuras pesquisas, onde se poderia ampliar o escopo investigativo a outros países e continentes com vistas à comparação e obtenção de conhecimento da realidade global. Uma breve busca por repositórios de dados científicos por país na base RE3DATA permite compreender a dimensão do problema a ser enfrentado pela enorme quantidade de repositórios indexados nela, sendo essa apenas uma fonte de outras possíveis.

Outras possibilidades de investigação que serviriam como extensão aos resultados encontrados estão no levantamento de dados quanto aos profissionais envolvidos nos projetos dos repositórios de dados científicos estudados, ou seja, formação profissional, composição de equipe(s), ferramentas utilizadas no fluxo de trabalho, etc. Seria igualmente possível explorar e comparar os projetos desses repositórios, as políticas institucionais de gestão e curadoria de dados científicos, o impacto dessas tecnologias no fazer científico diário, entre outros.

Uma pesquisa relevante estaria circundando a ideia da avaliação dos Princípios FAIR com vistas à análise e discussão de suas diretrizes. Seu resultado poderia confrontar ou corroborar sua adequação à realidade político-econômica do continente sul-americano, partindo do pressuposto que a aplicação desses princípios está condicionada a recursos humanos e financeiros, algumas vezes disponibilizados apenas por políticas públicas bem direcionadas e estruturadas.

Os resultados encontrados por meio do desenvolvimento dessa investigação podem ser úteis a grupos específicos de leitores, cujos interesses se apoiam na busca da compreensão dos fenômenos aqui abordados, como pesquisadores e professores. Sendo essa uma expectativa comum, posto que se trata de um trabalho científico, espera-se alcance externo às paredes de laboratórios de pesquisa, onde os achados sejam de valia a profissionais imbuídos a enfrentar os desafios emergentes no que concerne à gestão e curadoria de dados científicos.

Apesar da pesquisa tratar majoritariamente das infraestruturas do conhecimento, as quais abrangem repositórios (*software*) e seus conjuntos de dados científicos, além de normas/padrões internacionais como os Princípios FAIR, a organização do conhecimento não

se limita a recursos computacionais, pois depende também da capacidade cognitiva humana, de seu conhecimento tácito e pensamento crítico no desenvolvimento de melhores práticas.

À vista disso, entende-se que profissionais da informação devem buscar sua capacitação em dados, a começar pelo planejamento de projetos e políticas institucionais dirigidas à implementação de repositórios de dados científicos, passando pelo entendimento das divergentes necessidades entre comunidades, pelo conhecimento técnico computacional exigido a tais práticas, e idealmente, pela busca da padronização e manutenção desses serviços.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016. 320 p.

AMORIM, R. C. *et al*. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Univ Access Inf Soc**, Berlin, v. 16, p. 851-862, 2016. DOI: 10.1007/s10209-016-0475-y. Disponível em: https://link.springer.com/article/10.1007/s10209-016-0475-y. Acesso em: 02 jul. 2020.

BASKARADA, S.; KORONIOS, A. Unicorn data scientist: the rarest of breeds. **Program**, Northern Ireland, v. 51, n. 1, p. 65-74, 2017. DOI: 10.1108/PROG-07-2016-0053. Disponível em: https://www.emeraldinsight.com/doi/abs/10.1108/PROG-07-2016-0053. Acesso em: 20 jun. 2018.

BEAUJARDIÈRE, J. de la. NOAA environmental data management. **Journal of Map & Geography Libraries**, [*S. l.*], v. 12, n. 1, p. 5-27, 2016. DOI: 10.1080/15420353.2015.1087446. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/15420353.2015.1087446?tab=permissions&scroll=top. Acesso em: 13 jul. 2020.

BITBERRY SOFTWARE APS. **File.org**: dta. [*S. l.*], 2020. Disponível em: https://file.org/extension/dta. Acesso em: 21 fev. 2020.

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

BORGMAN, C. L; SCHARNHORST, A.; GOLSHAN, M. S. Digital data archives as knowledge infrastructures: mediating data sharing and reuse. **Journal of the Association for Information Science and Technology**, [*S. l.*], v. 70, n. 8, 2019. DOI: https://doi.org/10.1002/asi.24172. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.24172. Acesso em: 17 jun. 2020.

BROWNLEE, R. Research data and repository metadata: policy and technical issues at the University of Sydney Library. **Cataloging & Classification Quarterly**, [*S. l.*], v. 47, n. 3-4, p. 370-379, 2009. DOI: https://doi.org/10.1080/01639370802714182. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01639370802714182. Acesso em: 18 jun. 2020.

CRESWELL, J. W. **Research design**: qualitative, quantitative, and mixed methods approaches. 4. ed. Los Angeles: Sage, 2014. 340 p.

CROSAS, M. **The FAIR guiding principles**: implementation in Dataverse. Massachusetts, 2019. Disponível em: https://scholar.harvard.edu/mercecrosas/presentations/fair-guiding-principles-implementation-dataverse. Acesso em: 27 out. 2020.

CSIRO. **5-Star data rating tool**. [*S. l.*], 2017. *Software*. Disponível em: http://oznome.csiro.au/5star/?fbclid=IwAR2mZ21IMNInTxPYtX1Z2EqFdpof73vKSpBrCvJzBUvcvwHxRBmPcvUEfEc#page-top. Acesso em: 16 out. 2019.

DATAONE. **DataONE education module**: data management. [*S. l.*], 2012. *Powerpoint*. Disponível em: https://www.dataone.org/education-modules. Acesso em: 26 set. 2019.

DATAONE. **Software tools catalog**. [*S. l.*], [2020]. Disponível em: https://www.dataone.org/software_tools_catalog. Acesso em: 19 fev. 2020.

DATAVERSE. **Dataverse project**: about. [*S. l.*], [2020]. Disponível em: https://dataverse.org/about. Acesso em: 19 fev. 2020.

DKAN. **DKAN open data platform**. [*S. l.*], 2020. Disponível em: https://getdkan.org/. Acesso em: 06 ago. 2020.

DOORN, P.; TJALSMA, H. Introduction: archiving research data. **Arch Sci**, Netherlands, v. 7, p. 1-20, 2007. DOI: 10.1007/s10502-007-9054-6. Disponível em: https://link.springer.com/content/pdf/10.1007/s10502-007-9054-6.pdf. Acesso em: 21 abr. 2019.

EUROPEAN COMMISSION. **Turning FAIR into reality**: final report and action plan from the European Commission Expert Group on FAIR Data. Brussels, 2018. Disponível em: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Acesso em 27 out. 2020.

FIVESTARDATA. **5 Estrelas para dados abertos**. [*S. l.*], 2019. Disponível em: https://5stardata.info/pt-BR/. Acesso em: 16 set. 2019.

FORCE, M. M.; AULD, D. M. Data Citation Index: promoting attribution, use and discovery of research data. **Information Services & Use**, [*S. l.*], v. 34, n. 1-2, p. 97-98, 2014. DOI: 10.3233/ISU-140737. Disponível em: https://content.iospress.com/download/information-services-and-use/isu737?id=information-services-and-use%2Fisu737. Acesso em: 19 jun. 2020.

GARNETT, A. *et al*. Open metadata for research data discovery in Canada. **Journal of Library Metadata**, [S. l.], v. 17, n. 3-4, p. 201-217, 2017. DOI: https://doi.org/10.1080/19386389.2018.1443698. Disponível em: https://www.tandfonline.com/doi/full/10.1080/19386389.2018.1443698. Acesso em: 15 jul. 2020.

GO FAIR. **FAIR principles**. Germany; The Netherlands; Paris, 2019. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 4 set. 2019.

GÓMEZ, N. D.; MÉNDEZ, E.; HERNÁNDEZ-PÉREZ, T. Social sciences and humanities research data and metadata: a perspective from thematic data repositories. **El profesional de la información**, [*S. l.*], v. 25, n. 4, p. 545-555, 2016. DOI: http://dx.doi.org/10.3145/epi.2016.jul.04. Disponível em: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2016.jul.04/31589. Acesso em: 9 jul. 2020.

HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond, Washington: Microsoft Research, 2009.

HUNT, K. The challenges of integrating data literacy into the curriculum in an undergraduate institution. **IASSIST**, Denmark, v. 28, n. 2-3, p. 12-16, 2004. DOI: https://doi.org/10.29173/iq791. Disponível em: https://iassistquarterly.com/index.php/iassist/article/view/791. Acesso em: 21 ago. 2019.

IBICT. **Sistema para construção de repositórios institucionais digitais (DSpace)**. Rio de Janeiro; Brasília, 2019. Disponível em: http://www.ibict.br/tecnologias-para-informacao/DSpace. Acesso em: 08 out. 2019.

ILHARCO, F. Filosofia da Informação: alguns problemas fundadores. *In*: II Congresso Ibérico de Ciências da Comunicação, 2004, Portugal. **Anais** […]. Portugal, 2004. Disponível em: https://www.cccc2004.ubi.pt. Acesso em: 26 set. 2019.

KOLTAY, T. Data literacy: in search of a name and identity. **Journal of Documentation**, [*S. l.*], v. 71, n. 2, p. 401-415, 2015. DOI: 10.1108/JD-02-2014-0026. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-02-2014-0026/full/pdf?title=data-literacy-in-search-of-a-name-and-identity. Acesso em: 24 ago. 2019.

KOLTAY, T. Research 2.0 and research data services in academic and research libraries: priority issues. **Library Management**, [*S. l.*], v. 38, n. 6-7, p. 345-353, 2017. DOI: https://doi.org/10.1108/LM-11-2016-0082. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LM-11-2016-0082/full/pdf?title=research-20-and-research-data-services-in-academic-and-research-libraries-priority-issues. Acesso em: 23 out. 2019.

LEE, D. J.; STVILIA, B. Developing data identifier taxonomy. **Cataloging & Classification Quarterly**, [*S. l.*], v. 52, n. 3, p. 1-33, 2014. DOI: https://doi.org/10.1080/01639374.2014.880166. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01639374.2014.880166. Acesso em: 19 jun. 2020.

LIBRARY OF CONGRESS. **Preservation Metadata Maintenance Activity (PREMIS)**. Washington, DC, 2020. Disponível em: https://www.loc.gov/standards/premis/. Acesso em: 31 jul. 2020.

LTER; NCEAS. **Ecoinformatics.org**: Ecological Metadata Language. [*S. l.*], [2020]. Disponível em: http://ecoinformatics.org/tools.html. Acesso em: 23 mar. 2020.

MASHINGAIDZE, K.; BACKHOUSE, J. The relationships between definitions of big data, business intelligence and business analytics: a literature review. **Int. J. Business Information Systems**, [*S. l.*], v. 26, n. 4, 2017. DOI: 10.1504/IJBIS.2017.087749. Disponível em: https://www.inderscienceonline.com/doi/abs/10.1504/IJBIS.2017.087749. Acesso em: 25 jun. 2018.

MAYERNIK, M. S. Research data and metadata curation as institutional issues. **Journal of the Association for Information Science and Technology**, [*S. l.*], v. 67, n. 4, p. 973-993, 2015. DOI: 10.1002/asi.23425. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23425. Acesso em: 30 jun. 2020.

OPEN ARCHIVES. **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**. [*S. l.*], 2020. Disponível em: https://www.openarchives.org/pmh/. Acesso em: 03 ago. 2020.

OPEN GRAPH PROTOCOL. **Introduction**. [*S. l.*], 2020. Disponível em: https://ogp.me/. Acesso em: 21 fev. 2020.

RADIO, E. *et al*. Manifestations of metadata structures in research datasets and their ontic implications. **Journal of Library Metadata**, [*S. l.*], v. 17, n. 3-4, p. 161-182, 2017. DOI: https://doi.org/10.1080/19386389.2018.1439278. Disponível em: https://www.tandfonline.com/doi/full/10.1080/19386389.2018.1439278. Acesso em: 20 jul. 2020.

RESEARCH DATA ALLIANCE (RDA). **FAIR data maturity model**: specification and guidelines. RDA FAIR data maturity model Working Group, 2020. DOI: 10.15497/rda00045. Disponível em: https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf. Acesso em: 27 out. 2020.

RE3DATA. **RE3DATA.org**: about. [*S. l.*], 2019. Disponível em: https://www.re3data.org/about. Acesso em: 18 set. 2019.

ROCHA, L. L.; SALES, L. F.; SAYÃO, L. F. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso**, Salvador, v. 11, n. 3, p. 2-16, dez. 2017. DOI: http://dx.doi.org/10.9771/rpa.v11i3.24945. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/24945/15542. Acesso em: 20 set. 2018.

ROUSIDIS, D. *et al*. Metadata for big data: a preliminary investigation of metadata quality issues in research data repositories. **Information Services & Use**, [*S. l.*], v. 34, p. 279-286, 2014. DOI: 10.3233/ISU-140746. Disponível em: https://content.iospress.com/download/information-services-and-use/isu746?id=information-services-and-use%2Fisu746. Acesso em: 30 jun. 2020.

SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Conhecimento em Ação**, Rio de Janeiro, v.4, n. 1, p. 31-48, 2019. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337. Acesso em: 13 ago. 2020.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Persp. Ci. Inf.**, Belo Horizonte, v. 1, n. 1, p. 41-62, 1996. Disponível em: https://brapci.inf.br/_repositorio/2010/08/pdf_fd9fd572cc_0011621.pdf. Acesso em: 14 out. 2019.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. 90 p. Disponível em: http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf.

SCHOPFEL, J. *et al*. Open access to research data in electronic theses and dissertations: an overview. **Library Hi Tech**, [*S. l.*], v. 32, n. 4, 612-627, 2014. DOI: 10.1108/LHT-06-2014-0058. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LHT-06-2014-0058/full/pdf?title=open-access-to-research-data-in-electronic-theses-and-dissertations-an-overview. Acesso em: 25 jun. 2020.

SHARPENED PRODUCTIONS. **Fileinfo**: the files extension database. [*S. l.*], 2020. Disponível em: https://fileinfo.com/. Acesso em: 18 set. 2019.

STOREY, V. C.; SONG, I. Big data technologies and management: what conceptual modelling can do. **Data & Knowledge Engineering**, [*S. l.*], v. 108, p. 50–67, 2017. DOI: https://doi.org/10.1016/j.datak.2017.01.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169023X17300277. Acesso em: 25 jun. 2018.

SWAN, M. Philosophy of big data: expanding the human-data relation with Big Data science services. *In*: IEEE BigDataService, 2015, Redwood City, CA. **Anais** […]. Redwood City, CA, 2015. Disponível em: https://www.melanieswan.com/documents/Philosophy_of_Big_Data_SWAN.pdf. Acesso em: 21 set. 2019.

WAMBA, S. F. *et al*. How ‘big data’ can make big impact: findings from a systematic review and a longitudinal case study. **Int. J. Production Economics**, [*S. l.*], v. 165, p. 234-246, 2015. DOI: https://doi.org/10.1016/j.ijpe.2014.12.031. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925527314004253. Acesso em: 25 jun. 2018.

YU, H. H. The role of academic libraries in research data service (RDS) provision: opportunities and challenges. **The Electronic Library**, [S. l.], v. 35, n. 4, p. 783-797, 2017. DOI: https://doi.org/10.1108/EL-10-2016-0233. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/EL-10-2016-0233/full/html. Acesso em: 26 jul. 2020.

ZINS, C. Conceptual approaches for defining data, information, and knowledge. **Journal of the American Society for Information Science and Technology**, [*S. l.*], v. 58, n. 4, p. 479-493, 2007. DOI: https://doi.org/10.1002/asi.20508. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20508. Acesso em: 25 jun. 2018.