Artigo

em questão
E-ISSN 1808-5245

Repositórios de dados científicos na América do Sul: uma análise da conformidade com os Princípios FAIR

Marcello Mundim Rodrigues

Doutorando; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; marcellomundim@yahoo.com.br; ORCID: https://orcid.org/0000-0001-7945-6673

Guilherme Ataíde Dias

Doutor; Universidade Federal da Paraíba, João Pessoa, PB, Brasil; guilhermeataide@gmail.com; ORCID: https://orcid.org/0000-0001-6576-0017

Cíntia de Azevedo Lourenço

Doutora; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil; cintia.eci.ufmg@gmail.com; ORCID: https://orcid.org/0000-0002-2172-7300

Resumo: 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 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; os conjuntos de dados; os Princípios FAIR; e os repositórios digitais institucionais de dados científicos. O objetivo da pesquisa foi investigar a gestão e curadoria dos conjuntos de dados científicos disponibilizados nos repositórios digitais institucionais sul-americanos à luz dos Princípios FAIR. A proposta de investigação consistiu em uma pesquisa aplicada, de método qualitativo, exploratória, analítica, bibliográfica e documental. Fez-se levantamento dos repositórios de dados científicos no Registro de Repositórios de Dados de Pesquisa, o RE3DATA. A coleta dos dados foi feita nos repositórios selecionados. Utilizou-se da análise de conteúdo à concepção dos resultados de pesquisa. Os achados indicam que os programas responsáveis pelos repositórios investigados que servem à gestão e curadoria de dados científicos são o Morpho, o DSpace, e o Dataverse. 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: Dados científicos; Gestão de dados; Curadoria de dados; Princípios FAIR; Repositórios de dados

1 Introdução

Com o crescimento significativo do volume de dados ao longo dos anos de democratização tecnológica mundial, fase conhecida como globalização, aumentou-se 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* surgiu nesse contexto, o qual teve influência de outro cunhado por Weinberg em 1961, que definia o crescimento da pesquisa científica no mundo, a *Big Science* (BORGMAN, 2015).

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 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.

Nesse contexto, armazenar, descrever, organizar, tratar, e preservar dados e informação são práticas que podem ser consideradas complexas devido a inúmeros fatores que cercam o desenvolvimento de uma pesquisa científica e seus resultados. Costumeiramente, os dados não são publicados após o término de pesquisas, 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 ou fomento; uso de licenças; 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.

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.

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, a pergunta que se pretendeu responder é: qual a conformidade dos conjuntos de dados científicos arquivados em repositórios digitais institucionais sul-americanos com os Princípios FAIR? Os objetivos foram investigar conjuntos de dados científicos e respectivos repositórios digitais institucionais sul-americanos à luz dos Princípios FAIR.

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 Ciência da Informação (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.

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 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 iniciativas como a Research Data Alliance (RDA) e a Force11, poucos têm estudado o papel dos arquivos/repositórios de dados nas infraestruturas do conhecimento (BORGMAN; SCHARNHORST; GOLSHAN, 2019, p. 888-889, tradução nossa).

2 Referencial teórico

A pergunta a se fazer é "o que são dados?". Borgman (2015) responde que 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, tradução nossa).

Em resposta 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).

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).

O quarto paradigma da ciência se apoia nos anteriores, contudo, utiliza-se 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 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).

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[1] *apud* BASKARADA; KORONIOS, 2017, 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).

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 (NATIONAL SCIENCE BOARD, 2005[2] *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. Para Sayão e Sales (2015), "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). Amaral

(2016) entende que "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).

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).

Bibliotecários e profissionais da Ciência da Informação podem contribuir de forma vital com a curadoria de dados, a preservação, e 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, 2017, p. 66, tradução nossa).

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).

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

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, 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).

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).

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.

3.1 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 Instituições de Ensino Superior (IES) e agências de pesquisa sul-americanas recuperados pelo RE3DATA. 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.

3.1.1 Primeira etapa

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 Data Repository (Repositório de Dados de Levantamentos Biológicos); 2) Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) Dataverse Network; 3) Centro de Documentação e Acervo Digital da Pesquisa (CEDAP) Research Data Repository e; 4) Base de Dados Científicos da Universidade Federal do Paraná (UFPR).

Feito a primeira análise dos repositórios recuperados, identificaram-se inconsistências dos *link*s 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.

3.1.2 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; 6) International Center for Tropical Agriculture (CIAT) Dataverse (Colômbia); 7) Portal de Datos Abiertos de la Pontificia Universidad Católica (PUC) del Perú; 8) Repositorio de datos del Ministerio de Educación (MEC) del Perú; 9) Repositorio Institucional de la Universidad San Ignacio de Loyola (USIL) (Peru).

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

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 de a 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 viés (*bias*). 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 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.

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: Elaborado pelos autores.

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).

3.3 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 agosto de 2019 e janeiro de 2020.

3.3.1 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 (UFPR). A coleta foi registrada em planilhas, salvas nas categorias (colunas): Número do Depósito (número não fixo nos repositórios); URL ou URI do Depósito; e Identificador Persistente do Depósito.

3.3.2 Segunda etapa

Essa etapa repetiu a anterior e se utilizou do mesmo modelo de planilha para coletar dados dos demais repositórios sul-americanos selecionados.

3.3.3 Terceira etapa

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 2 e 3, e no Gráfico 1, localizados no próximo capítulo 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.

3.4 Procedimentos da análise dos dados

Para que se atingisse o resultado final dessa investigação, que corresponde ao objetivo de 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 1, 2, 3 e 4, 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.

3.4.1 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 (2017) 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), 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 Force11. O esquema das 5 Estrelas visa ajudar usuários a entender a maturidade de alguns dados ou serviços (CSIRO, 2017, tradução nossa).

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. A fim de satisfazer a natureza dessa investigação, os resultados numéricos das avaliações foram apresentados em texto (respostas dadas) nos Quadros 1 e 2.

3.4.2 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 à 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 3 e 4,

onde se percebem os critérios atendidos por repositório, seguidos dos motivos pelos quais o foram.

3.5 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: Universidade Federal de São Carlos (UFSCAR), Universidade Federal de São Paulo (UNIFESP), Universidade de São Paulo (USP), Universidade Estadual Paulista (UNESP), Instituto Tecnológico de Aeronáutica (ITA), Universidade Federal do ABC (UFABC), Universidade Estadual de Campinas (UNICAMP) e Empresa Brasileira de Pesquisa Agropecuária (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 se tornaria insatisfatório à investigação, pois quase todos os repositórios das instituições supracitadas foram estatisticamente eliminados, com exceção da EMBRAPA.

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.

A RE3DATA não possui todos os repositórios de dados científicos existentes na América do Sul indexados em sua base. 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.

4 Análise da conformidade com os Princípios FAIR

Esta seção apresenta a etapa de análise dos dados. Nela, encontram-se os resultados da análise de conteúdo dos objetos investigados à luz dos Princípios FAIR.

4.1 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 2 - 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: Elaborado pelos autores.

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 2, 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 1 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 1 - 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	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	Sim. g) Em serviço *Web* API padrão.	Sim. g) Em serviço *Web* API padrão.
3) Citável	Sim. d) Um identificador persistente *Web* (URI).	Sim. d) Um identificador persistente *Web* (URI).
4) Descrito	Parcialmente. d) Metadados especializados (ex.: *Darwin Core*, ISO 19115/19139, perfil dados científicos schema.org).	Parcialmente. c) Metadados básicos (ex.: Dublin Core).
5) Encontrável	Parcialmente. c) Em sistema de ampla comunidade ou de jurisdição.	Sim. d) Altamente ranqueado em índice de propósito geral (Google, Bing, etc.).
6) Carregável	Sim. c) Múltiplos formatos padrão.	Parcialmente. b) Um formato padrão, denotado por um tipo MIME.
7) Usável	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	Parcialmente. c) Etiquetas-padrão de comunidade (ex.: convenções CF, unidades UCUM).	Não. a) Códigos ou etiquetas de campo local.
9) Vinculado	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	Não. a) Nenhuma licença.	Sim. c) *Link* para uma licença padrão (ex.: *Creative Commons*).
11) Tratado	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.
12) Atualizado	Não. a) Conjunto de dados único.	Não. a) Conjunto de dados único.
13) Avaliável	Parcialmente. b) Declaração de linhagem em texto.	Não. a) Nenhuma informação sobre qualidade ou linhagem.
14) Confiável	Parcialmente. b) Estatística de uso disponível.	Não. a) Nenhuma informação sobre uso.
***Links* das avaliações**	https://oznome.csiro.au/5star?view=5e7a5b544d098386e957a18c	https://oznome.csiro.au/5star?view=5e826cc44d0983255557a210

Legenda: parcialmente = não atingiu nota máxima.
Fonte: Elaborado pelos autores.

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. Apesar de ser uma proposta interessante, 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 3 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 3 - 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: Elaborado pelos autores.

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 2, que serve como uma extensão do anterior, onde se percebe o número de questões com critérios atendidos por repositório.

Quadro 2 - 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	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	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.
3) Citável	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	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	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	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	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	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	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	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	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.	Sim. d) Repositório certificado.	Parcialmente. c) Repositório público ou institucional (ex.: CKAN, GitHub).
12) Atualizado	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	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	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: Elaborado pelos autores.

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.

4.2 Análise qualitativa da conformidade dos repositórios com os Princípios FAIR

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 3 apresenta os resultados das análises qualitativas dos conjuntos e repositórios brasileiros, seguindo o padrão comparativo estabelecido. 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. 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.

Quadro 3 - Conformidade dos repositórios brasileiros com os Princípios FAIR

FAIR	PPBio Data Repository *(*DataONE*)*	Base de Dados Científicos da UFPR (DSpace)
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.
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.
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.

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: Elaborado pelos autores.

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.

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. O Quadro 4 expõe a última parcela dos resultados, a qual diz respeito aos demais repositórios sul-americanos.

Quadro 4 - Conformidade dos demais repositórios sul-americanos com os Princípios FAIR

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)
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	tende 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.
A1	Atende totalmente: os conjuntos e seus (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.	Atende totalmente: os conjuntos e seus (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.	Atende totalmente: os conjuntos e seus (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.	Não atende: o identificador não disponibiliza acesso aos (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	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente	Atende totalmente: o HTTP é um protocolo de comunicação aberto, gratuito e universalmente	Não atende: não há protocolo de comunicação padronizado presente na configuração dos identificadores

	implementável.	implementável.	implementável.	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.
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, OpenAIRE, 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 Organization* (FAO) das Nações Unidas. 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.
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.

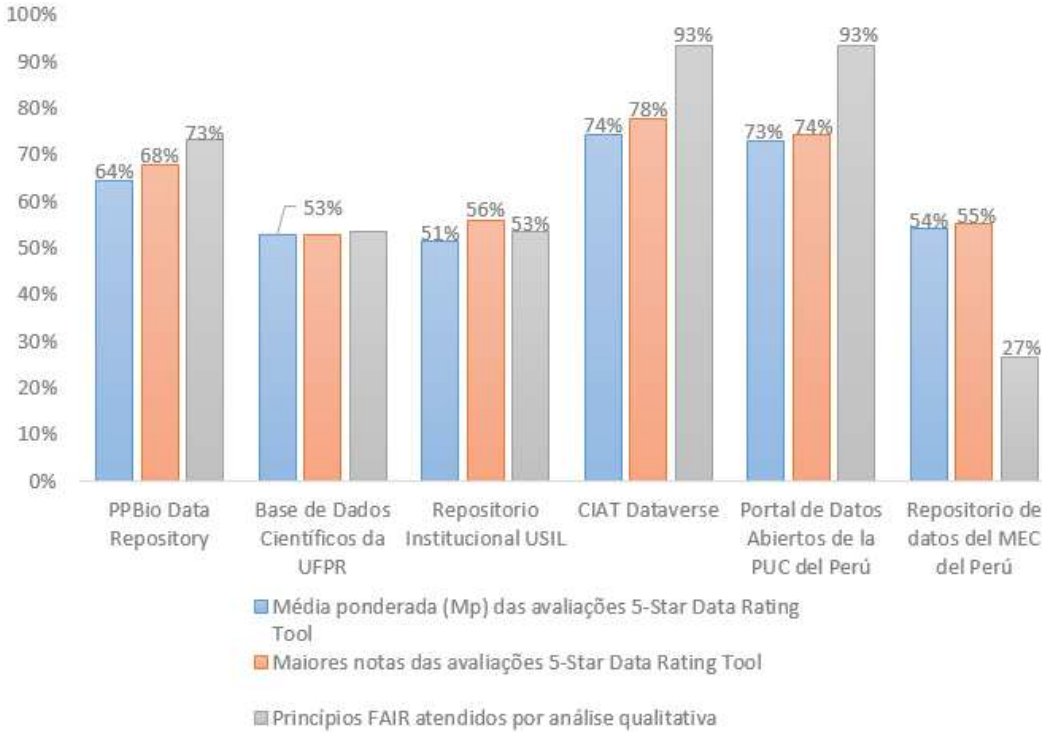
R1.1	Atende totalmente: há menção à licença *Open Database License* em texto e é possível acessá-la por *link*.	Atende totalmente: na aba *Terms*, é possível acessar o texto e *links* para as normas da comunidade e a licença *Creative Commons*.	Atende totalmente: na aba *Terms*, é possível acessar o 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.
Crit. Aten.	8 (52,8%)	14 (92,4%)	14 (92,4%)	4 (26,4%)

Fonte: Elaborado pelos autores.

Ambos repositórios da UFPR e do USIL atenderam totalmente a 8 critérios e a 1 parcialmente. 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.

Atendendo totalmente a 14 critérios, os repositórios do CIAT e da PUC Peru obtiveram resultados satisfatórios, especialmente o primeiro, 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: Elaborado pelos autores.

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, tradução nossa).

4.3 Discussão dos resultados

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.

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, 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.

5 Considerações finais

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.

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.

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

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

BASKARADA, S.; KORONIOS, A. Unicorn data scientist: the rarest of breeds. **Program: eletronic library and information systems**, Northern Ireland, v. 51, n. 1, p. 65-74, 2017.

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.

CROSAS, M. **The FAIR guiding principles**: implementation in Dataverse. Massachusetts: Harvard University, 2019.

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

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

EUROPEAN COMMISSION. **Turning FAIR into reality**: final report and action plan from the European Commission Expert Group on FAIR Data. Brussels, 2018.

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

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

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

RESEARCH DATA ALLIANCE (RDA). **FAIR data maturity model**: specification and guidelines. [*S. l.*]: RDA FAIR data maturity model Working Group, 2020.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.** Rio de Janeiro: CNEN, 2015.

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.

Scientific data repositories in South America: a FAIR compliance analysis

Abstract: 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; the datasets; the FAIR Principles; and the institutional digital repositories of scientific data. The research objective was to investigate the management and curation of scientific datasets deposited in south american institutional digital repositories in light of the FAIR Principles. The

investigation consisted of an applied, qualitative, exploratory, analytical, bibliographic and documentary research. The scientific data repositories were surveyed in the Registry of Research Data Repositories, better known as the RE3DATA. The data collection was made in the selected repositories. Content analysis was used to obtain the research results. The findings indicate that the software behind the investigated repositories which are fit to the management and curation of scientific data are Morpho, Dspace, and Dataverse. 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: Scientific data; Data management; Data curation; FAIR Principles; Data repositories

Recebido: 15/04/2021
Aceito: 29/07/2021

Declaração de autoria

Concepção e elaboração do estudo: Cíntia de Azevedo Lourenço, Guilherme Ataíde Dias; Marcello Mundim Rodrigues.

Coleta de dados: Marcello Mundim Rodrigues.

Análise e interpretação de dados: Marcello Mundim Rodrigues.

Redação: Marcello Mundim Rodrigues.

Revisão crítica do manuscrito: Cíntia de Azevedo Lourenço, Guilherme Ataíde Dias.

Como citar:

RODRIGUES, Marcello Mundim; DIAS, Guilherme Ataíde; LOURENÇO, Cíntia de Azevedo. Repositórios de dados científicos na América do Sul: uma análise da conformidade com os Princípios FAIR. **Em Questão**, Porto Alegre, v. 28, n. 2, e-113057, abr./jun. 2022. https://doi.org/10.19132/1808-5245282.113057

[1] PATIL, T.; DAVENPORT, D. Data scientist: the sexiest job of the 21st century. **Harvard Business Review**, Cambridge, MA, Out. 2012. Disponível em: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. Acesso em: 02 set. 2021. *Apud* Baskarada e Koronios (2017).

[2] NATIONAL SCIENCE BOARD. **Long-lived digital data collections**: enabling research and education in the 21st century. [*S. l.*]: National Science Foundation, 2005. Disponível em: https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf. Acesso em: 02 set. 2021. *Apud* Hey, Tansley e Tolle (2009).