PRINCÍPIOS FAIR APLICADOS À GESTÃO DE DADOS DE PESQUISA

LUANA FARIAS SALES
VIVIANE SANTOS DE OLIVEIRA VEIGA
PATRÍCIA HENNING
LUÍS FERNANDO SAYÃO
ORGANIZADORES

ibict

Princípios FAIR aplicados à gestão de dados de pesquisa

Esta publicação tem a cooperação da UNESCO no âmbito do projeto "Ampliação e Modernização das Ações do IBICT relacionadas às Atividades de Coleta, Armazenamento, Sistematização, Análise, Disseminação e Preservação de Dados e Informações Relativos à Ciência, Tecnologia e Inovação" (Prodoc 914BRZ2005). As indicações de nomes e a apresentação do material ao longo deste livro não implicam a manifestação de qualquer opinião por parte da UNESCO a respeito da condição jurídica de qualquer país, território, cidade, região ou de suas autoridades, tampouco da delimitação de suas fronteiras ou limites. As ideias e opiniões expressas nesta publicação são as dos autores e não refletem obrigatoriamente as da UNESCO nem comprometem a Organização.

PPGCI 50 ANOS
Programa de Pós-Graduação em Ciência da Informação | IBICT- UFRJ

COLEÇÃO PPGCI
50 ANOS

Princípios FAIR aplicados à gestão de dados de pesquisa

Luana Farias Sales
Viviane Santos de Oliveira Veiga
Patrícia Henning
Luís Fernando Sayão
organizadores

ibict
Instituto Brasileiro de Informação em Ciência e Tecnologia

Rio de Janeiro
2021

Capa: Fernanda Estevam
Ilustração: GK Vector (br.freepik.com)
Diagramação: Letícia Castro
Projeto Gráfico: Paulo César Castro
Normalização e catalogação: Selo Nyota

P957

Princípios FAIR aplicados à gestão de dados de pesquisa / Luana Farias Sales; Viviane Santos de Oliveira Veiga; Patrícia Henning; Luís Fernando Sayão (org.). – Rio de Janeiro: IBICT, 2021.

292p.-- (Coleção PPGCI 50 anos)

Inclui Bibliografia.
Disponível em: https://ridi.ibict.br/
ISBN 978-65-89167-24-2 (digital)
DOI: 10.22477/9786589167242

1. Dados de pesquisa. 2. Gestão de dados de pesquisa. 3. Dados abertos. I Sales, Luana Farias. II. Veiga, Viviane dos Santos de Oliveira. III. Henning, Patrícia. IV. Sayão, Luís Fernando. V. Título.

CDD 002:004

Projeto editorial em colaboração com o Programa de Educação Tutorial (PET) da Escola de Comunicação (ECO-UFRJ): Paulo César Castro (tutor) / aluno(a)s: Carolina Torres, Dandara Campello, João Maurício Maturana, Juliana Sorrenti, Kethury Santos, Lianne Henriques, Mariana da Paz, Ludmila Rancan, Moniqui Frazão, Robertha Braga, Sabrina Oliveira e Sara Maluf.

Programa de Pós-Graduação em Ciência da Informação (PPGCI), desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia, do Ministério da Ciência e Tecnologia e Inovação (IBICT/MCTI) em convênio com a Escola de Comunicação da Universidade Federal do Rio de Janeiro (ECO/UFRJ).

Rua Lauro Muller, 455 - 4° andar
Botafogo - Rio de Janeiro - RJ
http://www.ppgci.ufrj.br

Sumário

Seção 2

DADOS ACESSÍVEIS

Seção 3

DADOS INTEROPERÁVEIS

Seção 4

DADOS REUSÁVEIS

Seção 5

PRÁTICAS DE GESTÃO DE DADOS FAIR NO AMBITO DO GO FAIR BRASIL

A pesquisa que resulta nesta publicação obteve o fomento de
CNPq
FAPERJ
Capes

& com o apoio de
Unesco
IBICT
CENANCIN
UNIRIO
UFRJ
UFMG
Fiocruz

Para aqueles,
que como nós,
acreditam
que a
Ciência
torna o
mundo
melhor para
se viver.

"This easy access to information will transform the way we do science, the way we manage our businesses, the way we learn, and the way we play. It will both enrich and empower us and future generations"
(Gray, 1996)

Prefácio

Luiz Olavo Bonino da Silva Santos[1]

Desde o surgimento do movimento FAIR em janeiro de 2014 e da publicação de seus princípios em março de 2016, podemos observar um crescimento expressivo no interesse mundial em tratar melhor os dados e outros objetos digitais para que sejam mais facilmente encontráveis, acessíveis, interoperáveis e reusáveis. Os princípios foram definidos com o objetivo principal de expressarem um conjunto de comportamentos esperados para os objetos digitais a fim de torná-los mais suscetíveis à atuação de sistemas computacionais. Imediatamente observamos diversas iniciativas que buscavam opções de como criar implementações que seguissem os princípios. Porém, essas iniciativas ocorriam de forma independente, sem qualquer coordenação entre elas. Para evitar o desperdício de recursos e reduzir o risco de implementações incompatíveis, o que seria contrário às intenções originais do FAIR, os governos da Holanda e Alemanha inicialmente e posteriormente o governo da França, se comprometeram a financiar o estabelecimento de uma entidade que pudesse apoiar e coordenar o desenvolvimento de soluções seguindo os princípios FAIR. Com isso nasceu o movimento GO (*Global, Open*) FAIR, a partir da criação do Escritório Internacional de Apoio e Coordenação GO FAIR, GFISCO (GO FAIR *International Support and Coordination Office*), na sigla em inglês.

O GFISCO tem três sedes, em Leiden, Holanda, em Paris, França e em Hamburgo, Alemanha. Naturalmente, por ser financiado por três governos europeus, o foco do GFISCO deveria ser nas atividades e iniciativas europeias relacionadas ao FAIR no contexto da *European Open Science Cloud (*EOSC*)*. Porém os benefícios do EOSC só seriam maximizados se ele não se estabelecesse como um silo europeu mas como o braço europeu de um ambiente global, baseado nos princípios FAIR. Com isso, o GFISCO atua também para expandir o movimento GO FAIR internacionalmente. Os primeiros resultados desse esforço foram a criação de escritórios GO FAIR regionais no Brasil e nos Estados Unidos.

1 International Technology Advisor - GO FAIR International Support and Coordination Office, the Netherlands.

Este livro traz uma coleção de artigos que reportam trabalhos feitos no Brasil e apoiados pelo escritório GO FAIR Brasil. Como podemos ver pelo conteúdo, o movimento GO FAIR vem crescendo significativamente no país e cobre uma variedade de assuntos, indo da discussão do gerenciamento de dados e pesquisa à luz dos princípios FAIR até propostas de extensões em repositórios de dados para cumprirem com os requisitos dos princípios.

A pandemia do COVID-19 ao mesmo tempo que trouxe enormes problemas para todos os países do mundo, também vem servindo para deixar claro que temos ainda muito a fazer na área de gestão de dados. É cada vez mais claro que poderíamos ter um ganho de eficiência em quase todos os aspectos da resposta à pandemia, desde a identificação do crescimento no contágio até as análises dos resultados dos testes clínicas das vacinas, passando pela análise do processos biológicos relacionados à infecção do SARS-CoV2 e a eventual identificação de tratamentos adequados. Esse esforço global no combate à COVID-19 também se reflete neste volume, com trabalhos direita e indiretamente ligados a esse assunto.

O esforço para termos um mundo mais FAIR está apenas começando, mas se tivermos como indicativo o volume e a qualidade dos trabalhos apresentados neste livro, podemos ficar otimistas que esses objetivos serão efetivamente alcançados e teremos um ambiente onde objetos digitais interagem de forma mais eficiente e transparente.

Como citar com o DOI individual

SANTOS, Luiz Olavo Bonino da Silva. Prefácio. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 5-6. DOI: 10.22477/9786589167242.pref

Apresentação

Luana Farias Sales, Viviane Santos de Oliveira Veiga,
Patrícia Henning e Luís Fernando Sayão

O presente livro é uma tentativa de reunir iniciativas brasileiras teóricas e empíricas em torno da aplicação dos princípios FAIR e ser mais um instrumento de disseminação desses princípios no Brasil, especialmente no âmbito da pesquisa em Ciência da Informação e da Computação. Assim, após o prólogo, apresentamos um histórico de realizações da iniciativa GO FAIR no mundo e no Brasil, no periodo de 2017 a 2020. O livro está organizado considerando em um primeiro bloco, estudos que se desenrolam sob cada uma das categorias em que os princípios FAIR estão distribuídos e em um segundo bloco, relatos de experiências empíricas no âmbito do GO FAIR Brasil. Essas categorias não são mutuamente exclusivas de modo que alguns capítulos poderiam estar categorizados em mais de uma seção ou até em todas elas, afinal o livro é sobre aplicação dos princípios FAIR. No entanto, na tentativa de apresentar uma forma estruturada de organização, oferecemos essa organização aos nossos leitores.

Assim, no que tange à localização dos dados, o livro apresenta inicialmente o papel da identificação única de autor para que dados sejam encontrados, por meio da padronização dos nomes de autores, em seguida, nessa mesma seção optamos por categorizar os capítulos que se referiam à implementação dos princípios FAIR em repositórios por considerarmos esses uma ferramenta eficiente para tornar os dados localizados, seja através de APIs, seja através do OPI-MH. Ainda nessa seção, inserimos dois capítulos sobre experiências empíricas que embora não tenham acontecido no âmbito do GO FAIR Brasil, se enquadram perfeitamente dentro das iniciativas voltadas para facilitação da localização de dados de pesquisa.

No que tange à acessibilidade dos dados, a seção "Dados acessíveis" foi usada para abarcar os capítulos que tratam das experiências de acesso e uso de dados abertos em plataformas e repositórios de acesso aberto, como também uma interessante proposta teórica de aplicação de *linked data* em cadernos abertos de pesquisa.

Já na seção "dados interoperáveis", o livro traz a luz um estudo sobre padrões e outro sobre a noção de encapsulamento de dados para promoção de interopera-

bilidade de dados entre sistemas. Como experiência empírica, a seção traz ainda um estudo sobre interoperabilidade entre repositórios do centro de dados saúde do National Institute of Health (NIH).

No contexto do reuso de dados, reunimos três capítulos voltados para essa temática, o primeiro deles destaca a importância de incrementar o valor dos dados, ampliando sua visibilidade, bem como o potencial de reuso, para isso os autores apresentam um estudo em torno do significado dos termos *uso* e *reuso* e as condições que são estabelecidas para que o reuso seja efetivo. O segundo capítulo foca na necessidade de curadoria dos dados para que os dados sejam, além de FAIR, CARE, ampliando o foco da gestão de dados também para questões sociais, éticas e legais. De fato, os princípios FAIR abarcam apenas uma face da gestão de dados, que podem e devem ser complementadas com princípios que enfoquem também a governança em domínios específicos, promovendo assim o reuso consciente dos dados. Já o terceiro capítulo dessa seção aborda a criação de serviços de gestão de dados para promoção de seu reuso.

Finalizamos o livro então como uma seção, onde colocamos em destaque duas iniciativas relevantes no âmbito do GO FAIR Brasil: são elas a implementação da Rede GO FAIR Brasil Saúde-Enfermagem e por fim, o último capitulo vem apresentar uma riquíssima experiência surgida no auge do contexto pandêmico da COVID-19 com a finalidade de construir, de forma ágil, uma infraestrutura federada para uma rede de dados internacional, interoperável e distribuída, oferecendo suporte na busca por respostas, baseadas em evidências, sobre casos de surtos virais. A iniciativa internacional VODAN – sigla para *Virus Outbreak Data Network*, se estabeleceu no Brasil no contexto da Rede GO FAIR Saúde se valendo dos princípios FAIR para gestão dos dados e metadados coletados durante a pandemia.

Com esta reunião de capítulos, esperamos apoiar estudantes e pesquisadores que precisem se aventurar no amplo mundo de conhecimentos necessários para efetivar uma gestão de dados de pesquisa com qualidade, oferecendo um mix de conteúdos teóricos e empíricos sobre aplicação dos princípios FAIR.

Como citar com o DOI individual

SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando. Apresentação. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 7-8. DOI: 10.22477/9786589167242.apr.

Um panorama histórico da iniciativa GO FAIR: da Europa para o Brasil

Luana Sales[1], Viviane Veiga[2],
Patrícia Henning[3] e Luis Fernando Sayão[4]

Como tudo começou?

Após publicação do artigo *The FAIR Guiding Principles for scientific data management and stewardship*[5] na revista *Nature,* os princípios FAIR, um acrônimo para (*Findable, Accessible, Interoperable e Reusable*), obtiveram reconhecimento internacional assumindo o papel de referência mundial das boas práticas de gestão de dados.

Esses princípios desencadearam inquietação na comunidade acadêmico-científica quando foram inseridos na pauta das discussões do *Hight Level Expert Group on the European Open Science Cloud* (EOSC)[6] criado em 2016, pela Comissão Europeia. Esse grupo tinha o objetivo de apresentar recomendações que garantissem que ciência, negócios, governo e, eventualmente, a indústria, colhessem os melhores frutos da revolução do *big data*.

Desde então, diversas iniciativas foram surgindo gradativamente voltadas para o desenvolvimento de produtos e serviços que adotassem os princípios FAIR nas práticas da ciência. A iniciativa GO FAIR[7] foi uma delas, que despontou em 2016,

1 Luana Sales - doutora em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e Arquivo Nacional. <luanafsales@gmail.com>

2 Viviane Veiga – doutora em Ciências, Fundação Oswaldo Cruz, <viviane.veiga@icict.fiocruz.br>

3 Patrícia Henning – doutora em Ciências, Universidade Federal do Estado do Rio de Janeiro, <henningpatricia@gmail.com>

4 Luis Fernand Sayão – doutor em Ciência da Informação, Comissão Nacional de Energia Nuclear, <lsayao@cnen.gov.br>

5 https://www.nature.com/articles/sdata201618

6 https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

7 https://www.go-fair.org

de forma autônoma, com modelo de gestão *botton-up,* isto **é,** organizada de baixo para cima, pela própria comunidade, visando a disseminação e geração de recursos e serviços FAIR. A GO FAIR tem o intuito de garantir a devida reutilização dos dados em diferentes contextos, países e disciplinas, contribuindo para o compartilhamento e reuso de dados na geração de novos conhecimentos e para a reprodutibilidade da pesquisa. Essa iniciativa foi idealizada para atuar sob três pilares, **GO CHANGE**: investe em ações de divulgação procurando influenciar nas mudanças culturais e políticas que tornem os princípios FAIR referência na ciência; **GO TRAIN**: atua em treinamentos de diferentes níveis e naturezas, relacionados **à** aplicação dos princípios FAIR nas práticas da ciência; **GO BUILD**: impulsiona o desenvolvimento de infraestrutura técnica e operacional para suportar os dados FAIR.

Para expor melhor a atuação da iniciativa GO FAIR ao longo dos seus quatro anos de existência, este relato apresenta um panorama geral das principais realizações e conquistas alcançadas nesse período, dando destaque para a participação brasileira por meio da iniciativa GO FAIR Brasil.

2017: ano de expectativas e grandes desafios

O ano de 2017 foi de grande relevância para a iniciativa GO FAIR deslanchar e obter reconhecimento mundial. Sua política de atuação **foi motivada** e impulsionada pelas decisões da *Comissão Europeia*[8]**,** que passou a exigir que os resultados das pesquisas por ela financiadas, bem como seus respectivos dados, fossem disponíveis em acesso aberto e alinhados aos princípios FAIR. Além disso, **foi fortalecida** pelo grupo do (G7) composto por Canadá, França, Alemanha, Itália, Japão, Reino Unido e Estados Unidos da América, quando incluíram os princípios FAIR nas suas diretrizes de ciência aberta[9]**.** Por fim, **foi legitimada** quando os governos da Alemanha e da Holanda divulgaram o documento de posicionamento da EOSC se comprometendo a apoiar a iniciativa GO FAIR.[10]

As atividades começaram centradas no pilar *GO CHANGE*, direcionadas para o investimento na geração da cultura FAIR. Os resultados começaram a aparecer em publicações como a da revista *Nature,* intitulada *Don't let Europe's open-science dream drift*[11]. Esse artigo faz menção **à** iniciativa GO FAIR como aquela que deu o

8 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

9 https://www.dtls.nl/2017/10/04/g7-science-ministers-suggest-policy-guidelines-related-open--science/

10 https://www.government.nl/latest/news/2017/05/30/germany-and-the-netherlands-call-for--rapid-action-on-the-european-open-science-cloud

11 https://www.nature.com/articles/546451a

pontapé inicial em direção **à** nuvem científica europeia, fazendo com que as infraestruturas de dados concordassem em adotar protocolos que tornassem, pelo menos, alguns de seus dados FAIR. Outra publicação de igual relevância aparece no editorial da revista *Nature Genetics* intitulada *Data models to GO FAIR* relatando os aprendizados obtidos na realização de três *workshops* voltados para o gerenciamento de dados FAIR.[12] E ainda, nesse mesmo ano, é apresentado, em *preprint*, o artigo *A design framework and exemplar metrics for FAIRness*[13], que indica um conjunto básico de métricas semiquantitativas para a avaliação do nível de *FAIRness* dos dados.

Ainda em 2017 é lançada a primeira rede de implementação (RI) da GO FAIR, denominada *FAIR Metabolomics Network*.[14] O ano termina com a abertura do escritório internacional da iniciativa denominada GO FAIR *International Support and Coordination Office* (GFISCO)[15], apoiado pelos ministérios de pesquisa da Alemanha, Países Baixos e França, sediado em Leiden, na Holanda, dando início às suas atividades administrativas de apoio às RIs que estavam começando a se estruturar.

2018: ano de visibilidade internacional

O ano de 2018 começa administrativamente estruturado com a primeira reunião no GFISCO com o objetivo de: levantar o *status quo* das operações locais do escritório; identificar as ações em curso de relevância internacional; rever o andamento das adesões e lançamento das novas redes de implementação. Durante esse encontro foram estabelecidas ações no âmbito do pilar GO TRAIN – centradas em atividades de treinamentos de adoção dos princípios FAIR[16]. **É importante enfatizar que o GFISCO foi criado não apenas para apoiar** as atividades das redes de implementação, mas também orientar as RIs na elaboração de propostas de financiamento, bem como ajudar os *stakeholders* a "falar com uma só voz", fortalecendo e unificando o discurso a respeito dos princípios FAIR.[17]

A iniciativa GO FAIR começa a crescer e a expandir suas atividades pelo mundo a fora. Na França, surge por intermédio do Ministério do Ensino Superior, Pesquisa e Inovação, que promoveu um encontro com representantes de organizações de

12 https://www.nature.com/articles/ng.3910
13 https://www.biorxiv.org/content/10.1101/225490v3
14 https://www.go-fair.org/2017/03/16/metabolomics-implementation-network-launched-key-element-european-open-science-cloud
15 https://www.go-fair.org/go-fair-initiative/go-fair-offices/
16 https://zenodo.org/record/1168504#.WnwCi5POXOQ
17 https://www.go-fair.org/2018/02/20/report-meeting-potential-go-fair-implementation-networks

pesquisa e agências de fomento, apresentando os princípios FAIR e a iniciativa GO FAIR para a comunidade cientifica francesa.[18] Na Alemanha, apoiada pelo Ministério de Educação e Pesquisa, desponta na Universidade de Leibniz[19], com o primeiro *workshop* que apresenta os três pilares da iniciativa e faz uma análise das ações e esforços nacionais relacionados à gestão de dados de pesquisa naquele país.

No segundo semestre de 2018, começam a ser desenvolvidas ações focadas no pilar GO BILD. Uma delas foi um evento de repercussão internacional, na Universidade de Leinden, na Holanda, intitulado *Metadata for Machine* (M4M)[20]**.** Esse evento foi organizado pelas iniciativas GO FAIR e *Research Data Alliance* (RDA), trazendo reflexões sobre diferentes possibilidades de convergência de padrões de metadados e outras ferramentas de interoperabilidade para os serviços e dados FAIR. O evento foi dedicado à troca de conhecimento, sendo apresentadas iniciativas voltadas para os padrões de metadados e ontologias de diferentes tipos e infraestruturas, como as desenvolvidas pelas iniciativas *CLARIN*[21], *FAIR Data Point*[22], *CEDAR*[23] e *FAIRsharing*.[24] As discussões foram muito profícuas resultando em um projeto piloto em parceria com a principal agência de fomento holandesa, a ZonMW. Esse projeto foi publicado em *preprint*, intitulado The FAIR Funder pilot programme to make it easy for funders to require and for grantees to produce FAIR data[25]**,** que conta com a participação de representantes de instituições brasileiras.

Outra ação desenvolvida em 2018 foi a publicação de um estudo realizado por um grupo de especialistas que desenvolveram um conjunto básico de métricas semiquantitativas, com aplicabilidade universal, para a avaliação do nível de FAIRness dos dados. Esse artigo intitulado *A design framework and exemplar metrics for FAIRness*[26] encontra-se no servidor de preprint *bioRxiv.*

Nesse ano, a Iniciativa trabalha de forma atuante no âmbito do pilar GO CHANGE participando de eventos internacionais, buscando divulgar as suas ações e expandindo a cultura FAIR pelo mundo. Uma dessas participações foi por meio do

18 https://www.go-fair.org/wp-content/uploads/2018/03/The-First-French-GO-FAIR-Meeting-For-Future-INs.pd

19 https://www.go-fair.org/2018/10/08/on-the-road-to-fair

20 https://digitalscholarshipleiden.nl/articles/metadata-4-machines-help-you-find-and-reuse-relevant-research-data

21 https://www.clarin.eu/content/component-metadata

22 https://www.go-fair.org/how-to-go-fair/fair-data-point/

23 https://more.metadatacenter.org

24 https://fairsharing.org/

25 https://arxiv.org/abs/1902.11162v2

26 https://www.biorxiv.org/content/10.1101/225490v3

pôster denominado *The GO FAIR Approach: Building the EOSC Bottom-up – Based on Implementation Networks*[27], apresentado na conferência *Digital Infrastructure for Research* (DI4R)[28], realizada no Instituto Universitário de Lisboa (ISCTE). Esse evento levantou dúvidas relativas às questões de infraestruturas e desenvolvimento propostas nos projetos da União Europeia (EU).

As dúvidas continuavam persistindo demandando esclarecimentos nos eventos internacionais. Um deles ocorreu na Holanda, com a presença de dezoito países para discutir questões e equívocos relacionados aos princípios FAIR, tais como: O que é FAIR? O que não é FAIR? O que é GO FAIR? Como a iniciativa GO FAIR se relaciona com a EOSC e com a Internet de Dados e Serviços FAIR? E como a iniciativa GO FAIR está se juntando com os outros parceiros internacionais associados aos princípios FAIR?[29]

As Redes de Implementação (RI) não paravam de ser criadas. Nesse ano seis novas redes foram lançadas, entre elas: C2CAMP, *Genetic, Resarch Data Infraestructure*, *Economic and Social Sciences goING* FAIR, *Rare Diseases*, *Discovery.*

No final de 2018, foi criado o escritório *GO FAIR Brasil*[30] com atuação em todas as áreas do conhecimento. O seu primeiro encontro ocorreu no evento de comemoração dos 20 anos da *Scientific Eletronic Library Online* (SciELO), ocasião em que o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) assumiu a coordenação-geral do GO FAIR Brasil. E ainda nesse mesmo ano ocorreu o lançamento da primeira rede brasileira: a *GO FAIR Brasil Saúde,*[31] sob a responsabilidade do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict/Fiocruz), com a participação de diversas instituições das áreas de saúde pública, vigilância sanitária, informação e comunicação em saúde, história do patrimônio cultural das ciências e da saúde, oncologia, enfermagem e educação profissional em saúde.

2019: ano de consolidação

O ano de 2019 começa com o primeiro encontro anual das redes de implementação da iniciativa GO FAIR[32], em Leiden, na Holanda, com a presença de noventa profissionais das áreas de ciência da computação e ciência da informação, gestores de dados e áreas afins de diversos países da Europa, dos Estados Unidos e do Brasil.

27 https://www.go-fair.org/wp-content/uploads/2018/10/GO_FAIR_poster_DI4R.pdf
28 https://cetaf.eu/digital-infrastructures-research-di4r
29 https://www.go-fair.org/2018/11/07/go-fair-country-meeting-summary-and-outcome/
30 https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office/
31 https://portal.fiocruz.br/go-fair-brasil-saude
32 https://www.go-fair.org/2019/01/30/go-fair-implementation-networks-meeting

Esse evento teve o objetivo de aproximar a comunidade GO FAIR na identificação de possíveis formas de sincronização e convergência dos seus empreendimentos FAIR. A equipe do GFISCO expôs a política de governança da iniciativa GO FAIR e representantes dos EUA e do Brasil apresentaram o estágio de desenvolvimento da Iniciativa em seus países.

O ano de 2019 também foi contemplado com o lançamento de doze novas redes de implementação: BiodiFAIRse, *CO-OPERAS, FAIR StRePo, Chemistry, Novel Materials Discovery, Personal Health Train, Food Systems,* GlobAI I*ntegrated EArth Data,* GO Inte, Embassadors, *Data Stewardship Competence Centers e a GO FAIR África*[33]**,** com a apresentação do seu Plano de Implementação para o período 2019-2020, estabelecendo a conexão da África com a Internet de Serviços e Dados FAIR.[34]

O GFISCO promoveu o primeiro encontro para o desenvolvimento da Matriz de Implementação FAIR.[35] Nessa mesma época, aconteceu o *workshop* dedicado ao "GO CHANGE" com a participação de vinte e oito representantes de centros de competência em gerenciamento de dados de pesquisa da Áustria, Dinamarca, Holanda e Alemanha. Esse evento buscou estabelecer uma cultura acadêmica de dados de pesquisa "FAIR" tendo como resultado uma coletânea de artigos e documentos relacionados aos princípios FAIR.[36]

A participação em eventos continua sendo prioridade da iniciativa GO FAIR e de suas redes de implementação, que participaram da OSFair2019**,** sediada em Portugal.[37] Tanto a RI Discovery, quanto a RI CO-OPERAS realizaram sessões de *workshop* nesse evento. A sessão de pôsteres teve a participação da RI GO Inter e da iniciativa GO FAIR Brasil.[38] A Iniciativa GO FAIR esteve também presente na 14ª Plenária RDA, em Helsink[39]**,** apresentando a ferramenta *FAIR Convergence Matrix*, em desenvolvimento por um grupo de trabalho GO FAIR. Esse evento também contou com a participação brasileira apresentando o grupo recém-criado RDA Brasil, com apresentação do pôster intitulado *A Proposal of Machine-actionable Data Management Plan for Fiocruz.*[40]

33 https://www.go-fair.org/implementation-networks/overview/in-africa/

34 https://www.gzu.ac.zw/data-science-through-go-fair-in-africa-a-new-generation-internet-of-data-and-services/

35 https://www.go-fair.org/2019/06/19/fair-implementation-matrix-development-meeting/

36 https://www.zotero.org/groups/2345721/fair_data_resources/

37 https://www.opensciencefair.eu/

38 https://www.opensciencefair.eu/posters-2019/machine-actionable-data-management-plan-for-fiocruz

39 https://www.go-fair.org/2019/11/06/go-fair-at-rda-p14-in-helsinki/

40 https://www.rd-alliance.org/14th-plenary-call-posters

No âmbito do pilar GO TRAIN, os *workshops* tornaram-se uma prática recorrente. O 3º *workshop GOes FAIR*[41] aconteceu no *Institute GESIS Leibniz of Social Science*, em Colônia, na Alemanha, com a presença de várias instituições de pesquisa europeias que trataram de tópicos relacionados ao fornecimento de dados de pesquisa FAIR nos seus países.

Os dois últimos *workshops* do ano foram oferecidos pelo GFISCO, no ZBW - *Leibniz Information Center for Economics,* em Hamburgo, na Alemanha. O primeiro deles foi o *Semantic Interoperability of Metadata for Cross-Domain Research of the Future*.[42] O segundo foi *FAIR training and skills*[43], com a participação de administradores de dados de pesquisa da Alemanha, Holanda, Reino Unido, Suíça e Grécia.

Em 2019, ainda, o Coordenador-Geral da Iniciativa GO FAIR, Prof. Barend Mons, foi entrevistado pela revista cientifica brasileira LIINC em Revista[44], na edição especial sobre Dados de Pesquisa. Nessa entrevista foi relatada sua experiência de ter participado do *Hight Level Expert Group do EOSC* e sua vivência de pesquisador, na gestão de dados de pesquisa. O Prof. discorre, entre outras coisas, sobre os princípios FAIR, a Iniciativa GO FAIR, seu livro *Data Stewardship for Open Science: Implementing FAIR Principles*[45] e sua satisfação de ver como a iniciativa GO FAIR conseguiu ultrapassar as fronteiras da Europa chegando até ao Brasil.

Em novembro desse mesmo ano, ocorreu o primeiro seminário GO FAIR Brasil Saúde, no Instituto de Comunicação Científica e Tecnológica em Saúde – ICICT/ Fiocruz.[46] A mesa de abertura foi realizada pela vice-presidência da Fiocruz, a direção do ICICT e a coordenação da iniciativa GO FAIR Brasil Saúde. A palestra de abertura foi ministrada pelo representante da GO FAIR Internacional, na época, Dr. Luiz Olavo Bonino. Este seminário contou ainda com a participação da coordenação da GO FAIR Brasil, do representante brasileiro da GO FAIR internacional e dos coordenadores dos grupos de trabalho da GO FAIR Brasil Saúde.

Marca o término do ano de 2019 a reunião no GFISCO, com a presença de representantes da Alemanha, Bélgica, Brasil, Dinamarca, Grã-Bretanha, Itália, Holanda, Polônia, Suíça e EUA para o lançamento oficial da rede de implementação *Data Ste-*

41 https://www.go-fair.org/resources/go-fair-workshop-series/germany-goes-fair-workshops/
42 https://www.go-fair.org/2019/12/19/go-build-workshop-report/
43 https://www.go-fair.org/events/go-train-workshop
44 http://revista.ibict.br/liinc/article/view/5043/433
45 https://www.taylorfrancis.com/books/9781315380711
46 https://portal.fiocruz.br/noticia/seminario-da-rede-go-fair-brasil-saude-acontece-nos-dias-7-e-8-11

wardship Competency Centers (DSCC), contribuindo para a troca de conhecimentos entre as comunidades GO FAIR.[47]

2020: colhendo os frutos

O ano de 2020 começa com a 2ª reunião anual das redes de implementação da GO FAIR realizada na Alemanha, com a participação do representante da GO FAIR Brasil e de vinte e sete RIs, que se reuniram para explorar áreas de convergência cruzadas entre os domínios existentes.[48]

Uma edição especial da Revista *Data Intelligence*[49], publicada pela *MIT Press*, em 2020, traz uma compilação de artigos dedicados aos esforços da comunidade GO FAIR em torno das práticas emergentes dos serviços e princípios FAIR. O Brasil participou desta coletânea com o artigo GO *FAIR Brazil: A Challenge for Brazilian Data Science*.[50]

No mesmo ano aconteceu o curso eletivo *Introduction to FAIR Data Stewardship*[51], na *Hogeschool* na Holanda, desenvolvido sob o pilar GO TRAIN. Nesta mesma ocasião, o Prof. Barend Mons publicou na revista *Nature* um artigo intitulado *Invest 5% of research funds in ensuring data are reusable*[52] e Erik Schultes, da Iniciativa GO FAIR, publicou o artigo intitulado *A role for medical writersin overcoming commonly held misconceptions around FAIR data*[53]**,** na revista *Medical Writing Journal.*

Sete novas redes de implementação foram criadas em 2020: *FAIR Microbiome, Marine Data Centres, GO NANOFAB, Materials Cloud, AdvancedNano, GO UNI* e *Eco-Soc IN.*

Dois cursos de relevância internacional aconteceram neste ano. O primeiro foi o *Metatada for Machine*[54], patrocinado pela *Infrastructure Cooperation* (DeiC) da Dinamarca, em parceria com a Fundação GO FAIR; o segundo intitulado *Introduction to FAIR Data Stewardship*[55], ocorreu na *Hogeschool,* em Leiden, voltado para profissionais interessados em gestão de dados de pesquisa.

47 https://www.go-fair.org/implementation-networks/overview/dscc/
48 https://www.go-fair.org/2020/01/28/accelerating-convergence-in-2020/
49 http://www.data-intelligence-journal.org/p/issue/395
50 http://www.data-intelligence-journal.org/p/52/#:~:text=Today%2C%20GO%20FAIR%20Brazil%2DHealth,the%20process%20of%20adherence%20negotiation.
51 https://www.go-fair.org/events/introduction-to-fair-data-stewardship/
52 https://www.nature.com/articles/d41586-020-00505-7
53 https://journal.emwa.org/the-data-economy/a-role-for-medical-writers-in-overcoming-commonly-held-misconceptions-around-fair-data
54 https://www.go-fair.org/2020/07/08/m4m-for-the-danish-e-infrastructure-cooperation/
55 https://www.go-fair.org/2020/07/23/new-fair-data-stewardship-course-in-the-fall-of-2020/

A Rede de Implementação CO-OPERAS divulgou os relatórios dos cinco *workshops* organizados durante o ano de 2020 sobre o tema *"FAIR data for Social Sciences and Humanities"*, disponíveis no repositório Zenodo.[56]

O ano de 2020 é marcado por tristes surpresas e grandes preocupações em todo o mundo com a pandemia do novo *Coronavírus*. A Iniciativa GO FAIR foi particularmente tocada com o desafio de cumprir o seu papel de protagonista da gestor de dados alinhados aos princípios FAIR.

Na busca de soluções ao combate do novo *Coronavírus,* criou-se a rede de implementação da *Virus Outbreak Data Network* (VODAN)[57], focada na necessidade urgente de utilizar aprendizado de máquina e abordagens de inteligência artificial para descobrir padrões significativos para a gestão de dados em surtos epidêmicos. A VODAN é formada pela organização *Data Together* – composta pelas iniciativas CODATA, GO FAIR, RDA e WDS, que juntas formam uma infraestrutura federada. A VODAN tem por objetivo tornar os dados de surtos epidêmicos disponíveis para reutilização em novas pesquisas, monitoramento e outros fins, em condições bem definidas, respeitando a privacidade dos pacientes, conforme legislação vigente e apoiadas nos princípios FAIR.

No processo de divulgação da rede VODAN para a comunidade científica internacional, o Prof. Barend Mons publicou um artigo intitulado *The VODAN IN: support of a FAIR-based infrastructure for COVID-19*[58], na revista *European Journal of Human Genetics,* onde apresenta a rede VODAN e a infraestrutura criada para o enfrentamento da COVID-19. A partir desta iniciativa, focada para o enfrentamento da pandemia, outros países se associaram **à** rede, formando sub-redes.

A rede de implementação VODAN África[59] conta com a parceria de universidades, hospitais e ministérios da saúde de Uganda, Etiópia, Nigéria, Quênia, Tunísia e Zimbábue. O projeto teve ajuda internacional e o apoio da *Leiden University*, *GO FAIR Foundation, Tilburg University,* do *Europe External Program Africa* (EEPA) e da *Philips Foundation*. Por essa razão, a rede Vodan África está conseguindo atuar de forma estruturada destacando-se como referência internacional. A primeira instalação do grupo de trabalho *FAIR Data Point* para dados de COVID-19 foi feita na universidade internacional de Kampala, em Uganda, na África.[60]

56 https://www.go-fair.org/2020/08/28/co-operas-publishes-a-variety-of-workshop-reports--on-fairification-efforts-in-the-ssh/

57 https://www.go-fair.org/implementation-networks/overview/vodan/

58 https://www.nature.com/articles/s41431-020-0635-7

59 https://www.vodan-totafrica.info/

60 https://www.kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at-kampala-international-university_1595432235

A equipe *vodan África se junta com a Ásia* realizando juntos, em 2020, uma sessão técnica com o centro médico da Universidade de Leiden, para consulta intercontinental sobre o *fair Data Point*. As *Kampala International University, Stanford University, Leiden University* e a iniciativa go fair recebem financiamento do Google.org para a rede vodan. Esse financiamento foi destinado à implementação de padrões para o compartilhamento de dados e plataformas para modelagem de doenças em instituições de países na Uganda, Etiópia, Nigéria, Quênia, Tunísia e Zimbábue, visando o monitoramento e disseminação da covid-19.[61]

A vodan Brasil[62] tenta seguir o caminho da África com esforços voluntários de profissionais, pesquisadores e alunos da Fundação Oswaldo Cruz (Fiocruz), Universidade Federal do Rio de Janeiro (ufrj) e da Universidade Federal do Estado do Rio de janeiro (unirio) e com a participação do Hospital Universitário Gaffrée e Guinle, Hospital Municipal São José e Hospital Israelita Albert Einstein, conforme relatado no Capítulo 17 deste livro.

Ainda em 2020, ocorreu no Brasil o ii Seminário go fair Brasil Saúde em parceria com a unirio, intitulado Seminário Internacional sobre Gestão de Dados de Pesquisa em Saúde, com a participação de representantes da go fair Internacional e do Brasil.[63]

Quanto ao Brasil, os últimos acontecimentos do ano foram marcados (1) pelo lançamento da Rede go fair Brasil Saúde Enfermagem dentro das comemorações dos 130 anos da Escola de Enfermagem Alfredo Pinto, da Universidade Federal do Estado do Rio de Janeiro (unirio)[64], sob a responsabilidade do Programa de Pós-Graduação em Saúde e Tecnologia, no Espaço Hospitalar (*ppgsteh*), com relato mais aprofundado no Capítulo 16; (2) pela premiação brasileira de melhor pôster apresentado no encontro da 16ª Plenária do *Research Data Alliance* (rda), na Costa Rica, intitulado *vodan brazil - the Brazilian experience at the VIrus Outbreak Data Network65*; e (3) pela participação brasileira com a apresentação da vodan br no *International fair Convergence Symposiun 202066*, organizado pelas iniciativas

61 https://blog.google/outreach-initiatives/google-org/google-supports-covid-19-ai-and-data-analytics-projects

62 https://vodanbr.github.io/

63 http://www.unirio.br/prae/ppgsteh/noticias-1/seminario-internacional-sobre-gestao-de-dados-de-pesquisa-em-saude-1

64 https://www.go-fair.org/2020/09/12/launch-of-the-go-fair-brazil-health-nursing-network-on-september-22/

65 https://www.rd-alliance.org/rda-16th-plenary-meeting-poster-sessions

66 https://conference.codata.org/FAIRconvergence2020/

CODATA e GO FAIR. Esse evento proporcionou a criação de um fórum único para o avanço da convergência internacional entre diferentes domínios em torno dos princípios FAIR.

Considerações finais

Diante do panorama aqui apresentado através dos relatos dos principais acontecimentos que envolveram a Iniciativa GO FAIR durante o período de 2017 a 2020, percebe-se que a Iniciativa cresceu em âmbito nacional e internacional, se tornando uma das referências para a gestão de dados no mundo. Foram mais de vinte e sete redes de implementação criadas ao longo desses anos; diversos artigos foram publicados, vários encontros, eventos e treinamentos realizados em diferentes locais no mundo. O alinhamento das suas atividades com as políticas de ciência aberta da Comissão Europeia foi legitimado pelo EOSC, estando presente nos seus objetivos e atuação.

Percebeu-se ainda, que as ações que nos primeiros anos estavam voltadas para estudos teóricos/conceituais, definições de políticas, geração de cultura; treinamento e práticas de interoperabilidade relacionadas às aplicações do princípio FAIR, foram direcionadas para a rede de implementação VODAN. A pandemia da COVID-19 gerou pânico e preocupação mundial, impondo grandes desafios para a Iniciativa GO FAIR. A rede VODAN surgiu para ajudar no combate a esse vírus de alto contágio, assumindo o papel de desenvolvedor de infraestrutura para a gestão de dados de pacientes infectados pelo novo Coronavírus, alinhados aos princípios FAIR.

A participação brasileira sempre esteve presente junto à Iniciativa GO FAIR desde a época da criação do escritório da GO FAIR Brasil, no final de 2018. A GO FAIR Brasil tem se empenhando na busca de novas adesões, dentre as quais destacam-se a rede Humanidades, a rede Agro, a rede de Energia Nuclear, entre outras ainda em negociação. A rede GO FAIR Brasil Saúde é a mais estruturada até o momento, já contando com a sub-rede da área de Enfermagem. A Rede GO FAIR Brasil vem sendo representada nos eventos internacionais por meio do seu representante professor Dr. João Moreira, sediado na Holanda, na University of Twente.

É certo que os princípios FAIR expressam um caminho sem volta e que nunca estiveram tão presentes nas práticas mundiais de gestão de dados, como nos dias de hoje. Contudo, o que esperar de um futuro em que dados se colocam cada vez mais como insumo necessário para o desenvolvimento de novos conhecimentos, para a inovação e para a nossa sobrevivência como seres humanos e cidadãos? Muita coisa nos espera, especialmente em termos de avanços científicos e tecnológicos. No entanto, isso dependerá da forma como dados serão gerenciados. Neste sentido, é

fato que os princípios FAIR se apresentam como de fundamental importância para a promoção do compartilhamento de dados, de forma que sejam encontrados, acessados, interoperáveis e reusados. Assim, esperamos que com a organização deste livro possamos disseminar as pesquisas teóricas e empíricas realizadas no Brasil e incentivar nossos leitores e pesquisadores a aderirem ao movimento *GO FAIR*, participando de alguma forma da rede de implementação brasileira ou simplesmente aplicando os princípios FAIR na gestão de seus próprios dados de pesquisa.

Agradecimentos: Os organizadores deste livro gostariam de fazer um agradecimento especial para os professores da University of Twente, Dr. Luiz Olavo Bonino e Dr. João Moreira que juntos deram contribuições valiosas para a elaboração deste Prólogo bem como para o nosso conhecimento sobre os princípios FAIR.

Como citar com o DOI individual

SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando. Um panorama histórico da iniciativa GO FAIR: da Europa para o Brasil. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 9-22. DOI: 10.22477/9786589167242.cap1.

Seção 1

DADOS LOCALIZÁVEIS

FAIR PIDs: O papel da ORCID no fortalecimento dos Princípios FAIR

Paloma Marín-Arraiza[1], Ana Heredia[2]

1. Introdução

A ideia por trás de um sistema de identificadores persistentes (sistema PID) é oferecer uma referência duradoura a uma entidade (física, digital ou abstrata), por exemplo, um documento digital, site web, pessoa ou instituição. Alguns sistemas PID bem conhecidos são Archival Resource Key (ARK), Digital Object Identifier (DOI), Handle system, Persistent Uniform Resource Locator (PURL), Uniform Resource Name (URN) e Open Researcher and Contriburitor ID (ORCID iD), sendo esse último exclusivamente para pessoas.

Um PID possui uma série de metadados associados que são legíveis por máquinas, portanto, identificam o objeto e não a localização dele, como acontece com os URL (DAPPERT *et al.*, 2017). Um PID pode ser implementado seguindo o protocolo HTTP o que o torna acionável e permite dirigir ao leitor à página onde o recurso pode ser encontrado (*f*) (LÓPEZ-PELLICER *et al.*, 2016; VAN DE SOMPEL *et al.*, 2014).

No entanto, é importante ressaltar que a persistência está relacionada ao serviço oferecido pelo sistema e não ao identificador em si. Isto significa que uma entidade se compromete a manter o identificador resolúvel. O identificador leva os utilizadores aos serviços que garantem a referência (KUNZE, 2013). Por exemplo, os ARKs podem ser mantidos e resolvidos através do serviço EZID (Universidade da Califórnia); as DOIs são geridas pela International DOI Foundation e pelas suas agências de registo correspondentes, tais como Crossref e DataCite e centros de dados; os

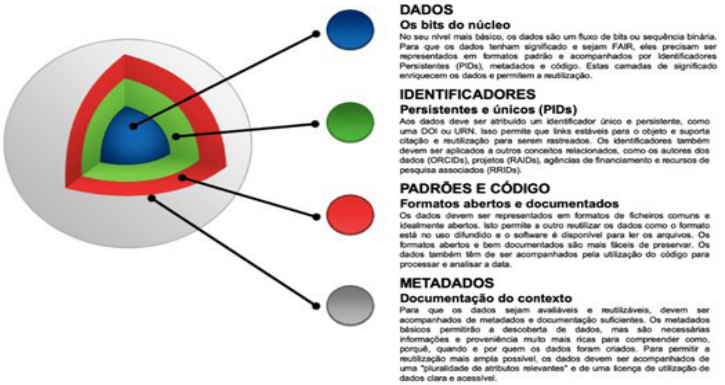
1 Dados da autora: Doutora em Ciência da Informação (Universidade Estadual Paulista), Mestre em Informação e Comunicação Científica (Universidade de Granada), Licenciada em Física (Universidade de Granada), ORCID, p.arraiza@orcid.org, https://orcid.org/0000-0001-7460-7794.
2 Dados da autora: Doutora em Ciências (Université Libre de Bruxelles), Mestre em Ciências Cognitivas e Neurociências (Université Paul Sabatier), Bacharel em Biologia (Universidade de Santa Úrsula), Consultora independente da informação, heredia.a@gmail.com, https://orcid.org/0000-0001-7862-8955.

Handles são gerenciados pela Corporation for National Research Initiatives (CNRI); e o sistema PURL foi desenvolvido pelo Online Computer Library Center (OCLC).

O uso de PIDs em arquivos e sistemas de informação de pesquisa está hoje em dia generalizado, e os PIDs são considerados uma parte crucial do processo de preservação. Por isso, várias instituições de pesquisa criaram centros de dados para registar os PIDs, a fim de preservar seus conteúdos e torná-los internacionalmente encontráveis e editáveis. O centro de dados (*data centre*) encarregado da emissão de um PID – por exemplo uma biblioteca de pesquisa – deve também realizar as tarefas de curadoria digital para garantir a manutenção dos metadados do recurso (JOHNSTON *et al.*, 2018).

De fato, as diretrizes atuais indicam o uso de PIDs, como é o caso do primeiro princípio FAIR: "Os (meta)dados são atribuídos com identificadores globalmente únicos e persistentes". O relatório "Turning FAIR into a reality" propõe um modelo de FAIR Data Objects (EUROPEAN COMMISSION. DIRECTORATE GENERAL FOR RESEARCH AND INNOVATION., 2018), cujas camadas consistem em metadados, padrões, identificadores e dados.

Figura 1 – Modelo do FAIR Data Object.



Fonte: European Comission. Directorate General for Research and Innovation (2018, p.38, tradução própria).

Para compreender o FAIR Data Object os autores expõem que:

> Os dados precisam ser acompanhados por Identificadores Persistentes (PIDs) e metadados básicos de descoberta para que possam ser encontrados, usados e citados de forma confiável. Além disso, os dados devem ser representados em formatos padronizados - e idealmente abertos - e ser ricamente documentados utilizando normas e vocabulários de meta-

> dados adotados pelas comunidades de pesquisa para permitir a interoperabilidade e a reutilização. O compartilhamento de código também é fundamental e deve incluir não apenas a fonte em si, mas também a documentação apropriada, incluindo declarações legíveis por máquinas sobre dependências e licenças. (EUROPEAN COMMISSION. DIRECTORATE GENERAL FOR RESEARCH AND INNOVATION, 2018, p.39, tradução própria).

Além da identificação, os PIDs são utilizados para agregar recursos. Os resultados da pesquisa com um PID são mais fáceis de rastrear, o que facilita as atividades de monitoramento da pesquisa. No entanto, como já foi mencionado, a persistência não é uma característica intrínseca de um PID, mas está relacionada com o serviço subjacente.

Nesse sentido, pode-se falar de "identificadores confiáveis" que são – além de persistentes – únicos, descritivos, interoperáveis e governados. O consórcio ODIN (ORCID e DataCite Interoperability Network) propôs as seguintes características para os identificadores confiáveis:

1) São únicos em escala mundial.
2) Resolvem como URI's HTTP persistentes com suporte para negociação de conteúdo.
3) Eles vêm com metadados que descrevem suas propriedades mais relevantes, incluindo um conjunto mínimo de elementos de metadados comuns.
4) São interligáveis.
5) São interoperáveis com outros identificadores através de elementos de metadados que descrevem a sua relação.
6) São administrados por uma organização que possui um modelo de negócio sustentável e uma massa crítica de organizações membros que concordaram com procedimentos e políticas comuns, tem um órgão dirigente e está comprometida com o uso de tecnologias abertas. (ODIN CONSORTIUM *et al.*, 2013, p. 19).

Além disso, os PIDs servem como mecanismos de crédito e atribuição, ao citar os resultados da pesquisa (MCMURRY *et al.*, 2017). Como afirmam Wilkinson *et al.* (2016), as infraestruturas científicas – por exemplo, repositórios, supercomputadores ou equipamentos físicos – também podem receber um PID.

2. FAIR PIDs e níveis de maturidade

Os PIDs podem ser internos —quando são utilizados dentro de uma organização; por exemplo, o identificador de um empregado ou estudante—, proprietários —quando são utilizados em um único sistema; por exemplo, o identificador

de autor da Scopus (Scopus Author ID)— ou abertos —quando apresentam uma interoperabilidade completa com outros sistemas e identificadores; por exemplo, um ORCID iD, um DOI ou um Uniform Research Identifier (URI). Esses últimos permitem o estabelecimento de conexões confiáveis entre recursos.

Ainda, Demeranville (2018) define FAIR PIDs adicionando mais características desejáveis aos sistemas de PID:

> **FAIR PIDs:** Estes PIDs não são apenas resolúveis, mas também podem ser usados para descobrir metadados abertos, interoperáveis e bem definidos contendo informações de proveniência de uma maneira previsível. São governados abertamente para o benefício da comunidade. Exemplo: Os DOIs são armazenados como URLs "https://doi.org/10.1/123", ou simplesmente "10.1/123". [...] Os DOIs são regidos pela Intenational DOI Fundation e os metadados anexos estão disponíveis sob uma licença CC0, o que significa que está aberta a todos. Os metadados contêm informações sobre a editora, a publicação, outros autores, financiamento e afiliação(ões), tudo isso ajuda a estabelecer a proveniência do item. Outros FAIR PIDs incluem identificadores arXiv, identificadores PubMed e PubMed Central e a maioria dos identificadores ISBN. (DEMERANVILLE, 2018).

Nesse sentido, é importante assinalar a maturidade da infraestrutura por trás desses PIDs. Podemos considerar a maturidade de uma infraestrutura quando é de uso comum na comunidade de pesquisa e entre disciplinas do conhecimento. Segundo a pesquisa desenvolvida por Ferguson *et al.* (2018) no marco do projeto FREYA, apenas as entidades "pesquisador", "publicação" e "dados" possuem na atualidade sistemas maduros de PIDs.

O seguinte quadro (quadro 1) mostra aqueles PIDs cuja infraestrutura possui um alto nível de maturidade.

Quadro 1 – Entidades, tipos de PIDs e sua maturidade.

Entidade de pesquisa	Tipos de PIDs usados	Maturidade da infraestrutura de PIDs
Publicação	DOI, Accession number, Handle, URN, Scopus EID, Web of Science UID, PMID, PMC, arXiv Identifier, BibCode, ISSN, ISBN, PURL	Madura
Pesquisador (ou acadêmico)	ORCID iDs, ISNI (também DAIs, VIAFs, arxivIDs, Open IDs, Researcher IDs, Scopus IDs)	Madura
Dados	DOI, Accession number, Handle, PURL, URN, ARK	Madura

Fonte: Adaptado de Ferguson et al. (2018, p. 9-10).

Os ORCID iDs fazem parte desta infraestrutura madura de pesquisa e também contribuem à FAIRificação dos dados de pesquisa como se descreve a seguir.

3. Findable e interoperable: o papel da ORCID

Os princípios FAIR guiam o processo de publicação de dados para torná-los *encontráveis* (findable – F), acessíveis (accessible – A), interoperáveis (interopera ble – I) e reutilizáveis (reusable – R).

O papel da ORCID no contexto dos princípios FAIR é entendido desde que o ORCID iD atua como padrão internacional na identificação persistente de autores. O quadro 2 apresenta esta contribuição para cada principio FAIR.

Quadro 2 – O papel da ORCID nos princípios FAIR

Princípio	Descrição[3]	Contribuição da ORCID
F1	Os (meta)dados recebem um identificador globalmente único e persistente	Fornecimento do ORCID iD como PID para "autor"/"criador" e "colaborador" .
F2	Os dados são descritos com metadados ricos[4]	Detalhe da informação de proveniência.
F3	Os metadados incluem clara e explicitamente o identificador dos dados que eles descrevem	Inclusão de PIDs em todas as entradas inseridas no registro ORCID.
F4	Os (meta)dados são registrados ou indexados em um recurso pesquisável	Disponibilização da ORCID Public API[5] para consultas. Publicação anual do arquivo[6] de dados públicos da ORCID.
I1	Os (meta)dados utilizam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento.	Reconhecimento do ORCID iD como norma ISO 27729:2012[7]. Utilização de padrões internacionais para a construção do registro (p.ex. CASRAI[8]).
I3	Os (meta)dados incluem referências qualificadas a outros (meta)dados	Apresentação utilizando HTTPS PIDs[9] para descobrir os metadados que descrevem o item vinculado

Fonte: Elaborado pelas autoras.

Desta forma, o uso de ORCID iDs (autenticados[10] se possível), contribui para o processo de FAIRificação. Ainda, isto, juntamente com o trabalho para melhorar a

3 Obtida e traduzida de https://www.go-fair.org/fair-principles/

4 Refere-se ao fato de ter atributos relevantes e com informação de proveniência

5 ORCID Public API: https://members.orcid.org/api/about-public-api

6 ORCID Public Data File 2020: https://doi.org/10.23640/07243.13066970.v1

7 ISSO 27729:2012. Information and documentation – International standard name identifier (ISNI) https://www.iso.org/standard/44292.html

8 CASRAI: https://casrai.org/

9 Identificadores contidos no registro ORCID: https://pub.orcid.org/v3.0/identifiers

10 Processo de obtenção de um ORCID iD autenticado: https://members.orcid.org/api/tutorial/get-orcid-id

qualidade e a integralidade dos metadados contidos nos registros ORCID, facilitará a ORCID se tornar uma fonte confiável de dados FAIR.

4. Notas finais

Este texto pretendeu apresentar alguns pontos sobre os PIDs e sua importância no contexto dos dados FAIR, bem como o papel da ORCID nos processos de FAIRificação dos dados.

ORCID, como organização sem fins lucrativos e fornecedora de infraestrutura aberta, continua se desenvolvendo e alinhando seu trabalho com a melhora da qualidade dos metadados e o apoio às comunidades de pesquisa. Os princípios FAIR baseiam também parte desse trabalho.

5. Referências

DAPPERT, Angela; FARQUHAR, Adam; KOTARSKI, Rachael; HEWLETT, Kirstie. Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. Data Science Journal, v. 16, p. 28, 15 jun. 2017. DOI 10.5334/dsj-2017-028. Disponível em: <http://datascience.codata.org/articles/10.5334/dsj-2017-028/>. Acesso em: 26 jun. 2020.

DEMERANVILLE, Tom. Blog: Building a Robust Infrastructure, One PID at a Time. , p. 0 Bytes, 2018. DOI 10.23640/07243.7008101.V1. Disponível em: <https://orcid.figshare.com/articles/Blog_Building_a_Robust_Infrastructure_One_PID_at_a_Time/7008101/1>. Acesso em: 1 nov. 2020.

EUROPEAN COMMISSION. DIRECTORATE GENERAL FOR RESEARCH AND INNOVATION. Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data. LU: Publications Office, 2018. Disponível em: <https://data.europa.eu/doi/10.2777/1524>. Acesso em: 1 nov. 2020.

FERGUSON, Christine; MCENTRYE, Jo; BUNAKOV, Vasily; LAMBERT, Simon; SANDT, Stephanie van der; KOTARSKI, Rachael; STEWART, Sarah; MACEWAN, Andrew; FENNER, Martin; CRUSE, Patricia; HORIK, René van; DOHNA, Tina; KOOP-JACOBSEN, Ketil; SCHINDLER, Uwe; MCCAFFERTY, Siobhan. D3.1 Survey Of Current Pid Services Landscape. 17 jul. 2018. DOI 10.5281/ZENODO.1324296. Disponível em: <https://zenodo.org/record/1324296>. Acesso em: 1 nov. 2020.

JOHNSTON, Lisa R; CARLSON, Jacob; HUDSON-VITALE, Cynthia; IMKER, Heidi; KOZLOWSKI, Wendy; OLENDORF, Robert; STEWART, Claire. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, v. 6, n. 1, p. 2198,

abr. 2018. DOI 10.7710/2162-3309.2198. Disponível em: <https://jlsc-pub.org/article/10.7710/2162-3309.2198/>. Acesso em: 9 mar. 2019.

KUNZE, J. The ARK Identifier Scheme. [*S. l.*]: California Digital Library, 2013. Disponível em: <https://tools.ietf.org/html/draft-kunze-ark-18>.

LÓPEZ-PELLICER, Francisco; BARRERA, Jesús; GONZÁLEZ, Julián; ZARAZAGA-SORIA, F.Javier; LÓPEZ, Emilio; ABAD, Paloma; RODRIGUEZ, Antonio F. El desafío de los identificadores persistentes y accionables. 2016. [*S. l.*: *s. n.*], 2016. Disponível em: <http://www.jiide.org/Jiide-theme/resources/docs/pdf/articulos/09_art_IAAA_IdentificadoresPersistentesAccionables.pdf>.

MCMURRY, Julie A.; JUTY, Nick; BLOMBERG, Niklas; BURDETT, Tony; CONLIN, Tom; CONTE, Nathalie; COURTOT, Mélanie; DECK, John; DUMONTIER, Michel; FELLOWS, Donal K.; GONZALEZ-BELTRAN, Alejandra; GORMANNS, Philipp; GRETHE, Jeffrey; HASTINGS, Janna; HÉRICHÉ, Jean-Karim; HERMJAKOB, Henning; ISON, Jon C.; JIMENEZ, Rafael C.; JUPP, Simon; … PARKINSON, Helen. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology, v. 15, n. 6, p. e2001414, 29 jun. 2017. DOI 10.1371/journal.pbio.2001414. Disponível em: <https://dx.plos.org/10.1371/journal.pbio.2001414>. Acesso em: 1 nov. 2020.

ODIN CONSORTIUM; ARYANI, Amir; BARTON, Amy J; BRASE, Jan; BROWN, Josh; DEMERANVILLE, Tom; HERTERICH, Patricia; MCAVOY, Lynne; PAGLIONE, Laura; RUIZ, Sergio; THORISSON, Gudmundur; VISION, Todd; ZIEDORN, Frauke. D4.2: Workflow for interoperability. [*S. l.*]: Figshare, 2015. DOI 10.6084/M9.FIGSHARE.1373669.V1. Disponível em: <https://figshare.com/articles/D4_2_Workflow_for_interoperability/1373669/1>. Acesso em: 26 jun. 2020.

VAN DE SOMPEL, Herbert; SANDERSON, Robert; SHANKAR, Harihar; KLEIN, Martin. Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. International Journal of Digital Curation, v. 9, n. 1, p. 331–342, jun. 2014. DOI 10.2218/ijdc.v9i1.320. Disponível em: <http://www.ijdc.net/article/view/9.1.331>. Acesso em: 9 mar. 2019.

WILKINSON, Mark D.; DUMONTIER, Michel; AALBERSBERG, IJsbrand Jan; APPLETON, Gabrielle; AXTON, Myles; BAAK, Arie; BLOMBERG, Niklas; BOITEN, Jan-Willem; DA SILVA SANTOS, Luiz Bonino; BOURNE, Philip E.; BOUWMAN, Jildau; BROOKES, Anthony J.; CLARK, Tim; CROSAS, Mercè; DILLO, Ingrid; DUMON, Olivier; EDMUNDS, Scott; EVELO, Chris T.; FINKERS, Richard; … MONS, Barend. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, v. 3, p. 160018, mar. 2016. Disponível em: <https://doi.org/10.1038/sdata.2016.18>.

Como citar com o DOI individual

MARÍN-ARRAIZA, Paloma; HEREDIA, Ana. FAIR PIDs: O papel da ORCID no fortalecimento dos Princípios FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 23 -30. DOI: 10.22477/9786589167242.cap2

Using the DATAVERSE project to move towards fair principles

Laura Vilela Rodrigues Rezende[1], Sonia Barbosa[2]

1. Introduction

Over time, the scientific context has been changing with the increase of movements in favor of the opening of Science and consequently the strengthening of sharing and collaboration. In this opening scenario, there are contextualized research data, which make it possible to confirm evidence from scientific studies. For this work, we will have as a basic premise the importance of sharing research data, which often have considerable potential for use, reuse and reinterpretation in different studies beyond the possibilities of reproduction. However, there are several challenges faced by the actors involved in opening and sharing scientific data. Among them, it is possible to list: difficulty of data interoperability; difficulty in locating scattered and disorganized data; high degradation rate of data links (supplementary), among others.

> There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly concerned with long-term data stewardship; and a data science community mining, integrating and analyzing new and existing data to advance discovery (wilkinson et al, 2016).

Faced with these challenges, several initiatives are underway to facilitate the opening and sharing of research data. In 2016, stakeholders representing academia, industry, funding agencies, and scholarly publishers designed and endorsed the

1 Doutora em Ciência da Informação. Universidade Federal de Goiás. lauravil.rr@gmail.com
2 ba Psychology & African American Studies/bsn. Harvard University. sbarbosa@g.harvard.edu

FAIR Data Principles, that may act as a guideline for those wishing to enhance the reusability of their data holdings (WILKINSON, 2016). For this purpose, the data must be Findable (Metadata and data should be easy to find for both humans and computers), Accessible (users need to know how they can access the data, possibly including authentication and authorization), Interoperable (the data need to interoperate with applications or workflows) and Reusable (metadata and data should be well-described so that they can be replicated and/or combined in different settings) (GOFAIR, 2020).

This paper aims to discuss how managing data using the Dataverse tool facilitates moving data towards FAIR principles by presenting five examples of data shared in the Harvard Dataverse (HD) repository. First, we will present the conceptual approach of the research data repository and the Dataverse Project; In the following topic the cases are presented and finally some conclusive analysis.

2. Research data repository: the dataverse project

The data management process consists of a set of practices that benefit current research project stakeholders (researcher, funding agencies, research institutions, among others) once it makes it possible to recover and share data for future research ensuring their integrity, reproducibility, and replicability. The process of management occurs at all phases of the research cycle, since the planning for data management, before the project begins; the documenting, organizing and securing data during the project; and finally archiving data after the research is completed. Despite the discipline-specific set of knowledge, practices, and skills related to research data lifecycle activities (collecting, creating, manipulating, analyzing, and sharing data), it is important to consider the research data repositories aiming to provide the sustainable infrastructure for the long term storage and access to research data.

The Research Data Repository is a database infrastructure set up to manage, share, access and archive well-described and well-documented research data. These databases may be specialized to aggregating disciplinary or more general data, collecting over larger knowledge areas.

The research data repository may provide all these resources listed in figure 1, improving the storing and sharing process. Besides general information and services, it must follow international standards related to technical aspects and metadata aiming to basically guarantee findability and interoperability. It must offer clear terms and conditions that meet legal requirements related to data protection, allowing use and reuse without unnecessary licensing conditions. These aspects provide achieving quality standards in the management and preservation of data.

Figure 1: The many planning aspects involved in the research data repositories

Available at: https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml

According to the Registry of Research Data Repositories (Re3data.org)[3], among several software available for data repositories, the most used are Dataverse[4], developed to store and share research data, D-Space[5], initially created for institutional repositories, CKAN[6], which was initially developed to promote the opening of government data, E-Prints[7], developed to research data. This study is part of an investigation carried out by the digital curation team responsible for the development of Dataverse, at Harvard University, which is why this software was chosen.

Based on the 2018 report - Open access to research data in Brazil: technological solutions (ROCHA, 2018) that the main reasons that make Dataverse the most used research data storage and sharing software, among other features, is that it has easily configurable resources for defining various types of environments and different characteristics for repositories, including different organizational hierarchies and management of policies for units or groups, various metadata and license schemes.

The Dataverse software was the brainchild of Dr. Gary King, Faculty Director of IQSS at Harvard University, to bring research data to the community and to make data FAIR, especially data that come with a scientific claim in related publications (KING, 2021).

A Dataverse repository is the software installation, which then hosts multi-

3 For more details, see: https://www.re3data.org/metrics/software
4 For more details, see: https://dataverse.org/
5 For more details, see: https://duraspace.org/dspace/
6 For more details, see: https://ckan.org/
7 For more details, see: https://www.eprints.org/uk/

ple virtual archives called dataverses. Each dataverse contains datasets (and may also contain other dataverses), and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data) (the dataverse project, 2020).

The Dataverse software is now in release 5.0 and continues to improve on its support of the fair principles, particularly in providing support for: persistent identifiers (at the dataset and file level), with url, and metadata registered to DataCite, customizable metadata (including support for multiple standards) exportable in numerous formats, versioning for datasets and files, deaccessioning of datasets (and versions of datasets), linked data support, data access and use terms, and file conversion to reusable formats[8]. It is important to note that while the Dataverse software helps to move data towards fair, data authors and collection managers must contribute to this goal by using appropriate community metadata and vocabularies standards.

In the next section, we present five collections of the hd repository to represent resources related to the fair principles served by the Dataverse software and their respective collection curation team.

3. Data shared in HD repository

Since the fair principles do not prioritize orienting issues related to data quality, but rather enhance their sharing, it must first be understood that making data aligned with these principles is a continuous process that requires, in addition to aligned technological aspects, considerable time, energy and expertise of those involved. The work of managing the collections´ data is essential in the process of alignment with the fair principles. With this in mind, the examples that will be presented below bring not only the technological resources implemented by default in the Dataverse software, but also some additional resources, policies, and workflows adopted that also increase and favor the data sharing process guided by the fair principles.

3.1 Methodological description

In order to carry out an analysis of the characteristics of some hd collections related to functions implemented by the software and best practices in data management actions, we sought to choose different segments that could represent different institutional and data generation contexts. One collection of: an organization/institution, a scientific journal, University Department, an individual re-

8 For more details, there is a Dataverse metadata standard page available at: https://guides.dataverse.org/en/latest/user/appendix.html

searcher, and a research group.

Regarding the analysis of the resources offered by the Dataverse software and the curation work carried out by the collection managers aligned with FAIR principles, the reference study chosen that presents the necessary interpretations and considerations was that of JACOBSEN, Annika et al. The authors presented the opinions of the original creators of the principles, supported by discussions of the experiences of pioneering FAIR implementers. They also pointed out the importance of presenting a common understanding around the original intentions of the guiding principles aiming to avoid divergence into non-interoperability.

3.2 The principle of "Findability"

The principle of findability, with its sub categories, are related to supporting users in their discovery process. It is considered the most fundamental of the FAIR principles, as globally unique and persistent identifiers are essential elements providing unambiguous identification of resources. In addition, this principle also contemplates facets of search, keywords and templates from the communities that facilitate capturing uniform and harmonized metadata.

The Dataverse citation resource, metadata tab of the dataset and files contain registered DOI and MD5 (UNF for tabular files) code. In addition, the software provides search facets, keywords and templates as resources related to discoverability. These are considered good practice in FAIR once the resource and its metadata are persistently linked and these identifiers may then successfully be used as the search term to discover its metadata record. Dataverse is also committed to using standard-compliant metadata to ensure that collections´ metadata can be easily mapped to standards schemas and exported format for preservation and interoperability.[9]

In the HD repository, the "Citation Metadata" element is the only required metadata block. This metadata element has five required fields for all datasets: "title and author name" are used to build the citation, and "e-mail contact, description, and subject" are used to enrich the dataset metadata and lend to the discoverability of content on the HD. Harvard Dataverse also supports DOIs at the file level. Collections created within the HD can utilize customization by selecting additional metadata elements to support their data. The table details the additional steps taken by the five examples in this study to enhance "findability" of their data, beyond the default software features:

9 For more details, there is a Dataverse metadata crosswalk available at: https://docs.google.com/spreadsheets/d/10Luzti7svVTVKTA-px27oq3RxCUM-QbiTkm8iMd5C54/edit#gid=222839033

Table 1: "FINDABLE" FAIR Principle in selected HD Collections[10]

	F1: unique and persistent identifier	F2: data are described with rich metadata	F3: metadata clearly and explicitly include the identifier of the data it describes	F4: (meta) data are registered or indexed in searchable resource
Organization: The International Food Policy Research Institute (IFPRI)[11]	YES (software implemented at dataset and file level)	Uses multiple metadata blocks; Uses optional metadata fields; Links to keyword and topic classification standard vocabulary; "widget" feature; "file tags"; additional metadata "terms" .	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	YES (Software implemented)
Journal: American Journal of Political Science (AJPS)[12]		Uses multiple metadata blocks, including "journal metadata block" and "related publication" metadata field; uses optional metadata fields; "file tags"; "widget" feature;	Always - "Related publication" metadata field to connect to journal article (bidirectional link)	
Department: Harvard University Department of Government[13]		Uses multiple metadata blocks; uses optional metadata fields; "file tags"; "widget" feature; uses additional metadata terms;	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	
Researcher: Gary King[14]		Uses multiple metadata blocks; uses optional metadata fields; "widget" feature; "file tags".	When available - "Related publication" metadata field to connect to journal articles (bidirectional link). Links to replication software utilized by the dataset.	
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)[15]		Uses multiple metadata blocks; uses optional metadata fields; uses additional metadata "terms"; "file tags" .	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	

source: self elaboration (2020)

10 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features

11 The IFPRI dataverse is available at: https://dataverse.harvard.edu/dataverse/IFPRI

12 The AJPS dataverse is available at: https://dataverse.harvard.edu/dataverse/ajps

13 The Harvard University Department of Government dataverse is available at: https://dataverse.harvard.edu/dataverse/GovDept

14 Gary King dataverse is available at: https://gking.harvard.edu/data

15 The SIIL dataverse is available at: https://dataverse.harvard.edu/dataverse/SIIL

3.3 The principle of "Accessibility"

The Dataverse software supports the principle of Accessibility with support for full dataset citations, DOIs with URLs, and metadata registered to Data Cite[16]. Citation and discoverable metadata are available using several standards, including schema.org, Dublin Core, and DDI. Data files can be restricted (authentication/ authorization required) or open for access. There is a Terms landing page with metadata for usage information. There is a citation for each data file, with a DOI and URL for each file. Downloads of the metadata include machine-actionable dataset landing pages with meta-tags for citation metadata. The Deaccession feature allows removal of a dataset and leaves a "tombstone" citation page which is findable and citable; metadata includes reason for "deaccessioning" / "Versioning." There is also Support for the web protocol HTTP (W3C); the data transfer protocol with mirroring, incremental backups, and file copies between systems: Rsync over ssh (GNU GPL); RESTful (Representation State Transfer) API; Authentication API Tokens; Authorization service.

Harvard Dataverse uses CCo[17] license by default, and allows depositors to opt out and use their license of choice. Depositors can choose whether their data are open or restricted for access, but in the latter case they must enable the "request access" feature for data requestors, or provide terms describing how users can request access to restricted content, or if content is embargoed for a period of time. Following DataCite standards, all metadata for datasets are visible and discoverable, even if files are not immediately downloadable for access. The examples in the table below include open data, embargoed content, and content that requires additional contact with the data owners for access. The table details the additional steps taken by the five examples in this study to enhance "accessibility" of their data, beyond the default software features:

16 For more details about Data Cite: https://datacite.org/

17 The HD licenses and terms of use are described in: https://dataverse.org/best-practices/harvard-dataverse-general-terms-use

Table 2: "ACCESSIBLE" FAIR Principle in selected HD Collections[18]

	A1: (meta)data are retrievable by their identifier using standardized communications protocol	sub-principle A1.1: the protocol is open, free and universally implementable	sub-principle A1.2: the protocol allows for an authentication and authorization procedure, where necessary	A2: metadata are accessible, even when the data are no longer available
Organization: The International Food Policy Research Institute (IFPRI)	YES (software implemented)		"File restriction" feature; "request access" feature.	YES (Software implemented)
Journal: American Journal of Political Science (AJPS)			Restricted content provides copyright info and access information provided.	
Department: Harvard University Department of Government			"File restriction" feature, with embargo.	
Researcher: Gary King			Restricted content provides copyright info and access information provided.	
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)			"File restriction" feature; "request access" feature.	

source: self elaboration (2020)

3.4 The principle of "Interoperability"

The Dataverse software supports the principle of "Interoperable" by supporting variable metadata for tabular data files, using DDI standard, Machine-actionable Variable description from DDI, and summary statistics in DDI automatically calculated upon data upload.

Harvard Dataverse integrated the Data Explorer[19] tool developed by Scholars Portal. Data Explorer is a GUI (Graphical User Interface) which lists the variables in a tabular data file allowing searching, charting and cross tabulation analysis. Every example in our table below utilizes the tabular data functionality where possible. The HD also uses the File Previewer tool, a set of tools that display the content of files - including audio, html, annotations, images, PDF, text, video, tabular data, and spreadsheets - allowing them to be viewed without downloading. The table details

18 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features

19 This feature is described here: https://guides.dataverse.org/en/latest/admin/external-tools.html

the additional steps taken by the five examples in this study to enhance "interoperability" of their data, beyond the default software features:

Table 3: "INTEROPERABLE" FAIR Principle in selected HD Collections[20]

	I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation		I2: (meta) data use vocabularies that follow fair principles *complete	I3: (meta)data include qualified references to other (meta)data
Organization: The International Food Policy Research Institute (IFPRI)	YES (software implemented) and integrated tools (Data explorer, File Previewer)	Of 10k files, 7k are tabular files	"keywords," & "Topic Classification" controlled vocabulary w/links to standards http://aims.fao.org/	When available - "Related publication" metadata field to connect to journal articles (bidirectional)
Journal: American Journal of Political Science (AJPS)		Of 8500k files, 2200 are tabular files *note this is a replication data journal so each dataset normally contains one data file, and one code file, and one readme file	Uses multiple metadata blocks, including "journal metadata block" and "related publication" metadata field; uses optional metadata fields; "file tags" ; "widget" feature; uses Center for Open Science "Open Materials and Open Data" badges.[21]	Always - "Related publication" metadata field to connect to journal articles (bidirectional)
Department: Harvard University Department of Government		Of 1350 files, 478 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; "file tags"; "widget" feature; uses additional metadata terms;	When available - "Related publication" metadata field to connect to journal articles (bidirectional)
Researcher: Gary King		Of 1870 files, 563 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; "widget" feature; "file tags".	When available - "Related publication" metadata field to connect to journal articles (bidirectional)
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)		of 1280 files, 287 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; uses additional metadata "terms"; "file tags" .	Links to associated manuscripts where possible; uses additional metadata "file tags"

source: self elaboration (2020)

20 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features

21 This feature is to acknowledge open practice of the dataset: https://osf.io/tvyxz/wiki/home/ ;

3.5 The principle of "Reusability"

The Dataverse software supports the principle of Reusability by supporting the integration of Make Data Count[22]. Citation and discoverable metadata using DataCite, schema.org, Dublin Core, DDI standards. Additional metadata support, including domain specific. Terms with license usage or data use agreement. PROV metadata (provenance). Domain relevant file download standards. Variable metadata for tabular data files using DDI standards, machine actionable variable descriptions from DDI, summary statistics in DDI, automatically calculated upon data upload.

Harvard Dataverse makes use of the Make Data Count integration. Provenance information is requested at the dataset level. The use of the Data Explorer tool allows for analysis and visualization of tabular data files. The table details the additional steps taken by the five examples in this study to enhance Reusability of their data, beyond the default software features:

Table 4: "REUSABLE" FAIR Principle in selected HD Collections

	r1: (meta) data are richly described with a plurality of accurate and relevant attributes	r1.1: metadata are released with a clear and accessible data usage license	r1.2: (Meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Organization: The International Food Policy Research Institute (IFPRI)	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including:more than one "keyword," more than one "topic classifications," and "social science" and "geospatial" metadata, " in addition, linking to standard agricultural standard vocabulary.	Public facing additional License and term of access / Data sharing agreement / Open Access and Open Data Policy / Donors policy 7	**Citation metadata block**, in addition: "grant information; distributor information; dates metadata, "contributors", "software," "series," "related publication and datasets," "data collectors," "data source." **Geospatial metadata block**: coverage country/nation; coverage state/province; coverage city; unit. **Social Science and Humanities metadata block**: "universe;" "unit of analysis;" "sampling procedure;" "collection mode;" "type of research instrument." **Use of templates** for consistency in required metadata fields and formatting of information in such fields.	Metadata Blocks used: Citation; Geospatial; Social Science; Additional metadata: Dataverse and Datasets description; Summary of collection content; Links to relevant documentation; search facets; templates

22 For more details see: https://makedatacount.org/

Table 4: "REUSABLE" FAIR Principle in selected HD Collections

	r1: (meta) data are richly described with a plurality of accurate and relevant attributes	r1.1: metadata are released with a clear and accessible data usage license	r1.2: (Meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Journal: American Journal of Political Science (AJPS)	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: more than one "keyword;" reproducibility verification workflow	Public facing verification police document, CC0 by default8; open to allow authors to use other licenses as needed; restricted content is clearly labeled with copyright and access information	**Citation metadata block**, in addition: "dates" metadata, "related publication and datasets; **Social Science metadata block; Geospatial metadata block: Journal Metadata Block**; Use of templates for consistency in required metadata fields and formatting of information in such fields.	Metadata Blocks used: Citation; Journal; Geospatial; Social Science; Additional metadata: Dataverse and Datasets description; Summary of collection content; Links to relevant documentation; search facets; templates
Department: Harvard University Department of Government	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: more than one "keyword," "topic classifications, "kind of data","related materials"; "related dataset".	Files embargoed with date of release; resolves to CC0 once released; Data PASS Terms standard 1.09	Citation metadata block; in addition: producer and distributor information; "dates" metadata, "related publication and datasets, Geospatial metadata block: "geographic coverage"; Social Science and Humanities metadata block: "unit of analysis;" templates	Metadata Blocks used: Citation; Social Science; Geospatial; Additional metadata: Dataverse and Datasets description; Summary of collection content; search facets; templates; links to relevant documentation
Researcher: Gary King	Use of dataverse templates to ensure consistency of metadata; lengthy dataset description, use of additional citation metadata fields, including: more than one "keyword," more than one "topic classification."	CC0 by default; restricted content is clearly labeled with copyright and access information	Citation metadata block; Use of templates	Metadata block used: Citation. Additional metadata: Dataverse and Datasets descriptions; Summary of collection content; search facets; templates

Table 4: "REUSABLE" FAIR Principle in selected HD Collections

	r1: (meta) data are richly described with a plurality of accurate and relevant attributes	r1.1: metadata are released with a clear and accessible data usage license	r1.2: (Meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)	Metadata verification via established workflows; use of dataverse templates to ensure consistency of included metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: "keyword," and Social Science and Humanities, Geospatial,Life Sciences, and Journal metadata blocks	Default CC0 waived with SIIL terms of use clearly defined	Citation metadata block; Use of templates	Metadata block used: Citation, Social Science and Humanities, Geospatial,Life Sciences, and Journal. Additional metadata: Dataverse and Datasets descriptions; summary of collection content; search facets; templates

source: self elaboration (2020)

4. Conclusions

It is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability (wilkinson, 2016). Important to note is that while the software provides functionality to help move data towards fair, data authors, collection managers and curators may contribute to this effort by utilizing the features provided and using best practices when sharing their data. The five examples in this paper detail the use of features to move their collections towards fair, and demonstrate not only the differences between the collections utilization of workflows and tools, but the impact of such use on fair.

The ifpri and feed the Future dataverses demonstrate the effectiveness of planned workflows and use of templates to guide large teams of curators, and organization level terms of access. ifpri, in linking to their standards, demonstrates the use of optional, but encouraged, standards and features. Both utilize a team of curators and data managers, and make use of the Featured Dataverses option to highlight their collections, and use additional metadata blocks offered by the software to describe their data, and extensive searchable metadata facets to improve the

Findability of their datasets. They also make use of multiple Dataverse templates prepopulated with the appropriate terms of use for each collection, saving curators and data managers the time needed to complete this information for each dataset. Of the five collections, IFRPI and Feed the Future utilize the request access feature for restricted files which allows them access control. The request access works in alignment with the terms of access to fulfil one sub principle of "accessibility."

The AJPS journal and Gary King dataverses are cases that support replication verification in data sharing. The AJPS dataverse utilizes the "submit for review" workflow to verify reproducibility of data prior to data publishing. This process is supported by the National Academies of Sciences (2020) statement that, "journals should consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible." In addition, AJPS demonstrates the desired level of curation and workflow to achieve and allow others to reproduce or replicate their results, as recommended by NAP that, "Journalists should report on scientific results with as much context and nuance as the medium allows." The Harvard Dataverse is a medium that allows rich reporting of scientific results via the Dataverse features and best practices guidelines. The journal's use of Open Badges is an additional acknowledgment to the authors for depositing content that is verified reusable. Gary King's dataverse, self-curated with numerous replication data supported by author confirmed reproducibility verification, are designated "replication" datasets that include data, code, documentation, and links to software[23] that allow maximum reuse of the data. Replicability of data, as demonstrated by the two collections, is aligned with the sub principle of Reusability associated with detailed provenance.

The Harvard Department of Government Dissertation Dataverse was selected as a unique case supporting the early engagement of graduate students in data sharing. All students in this department are required to share their dissertation data within Harvard Dataverse to fulfill their graduation requirement. The space was designed by the Harvard Curation team utilizing templates with prefilled metadata fields, and instructions for data deposits. The terms of use section of the template is prefilled with a 5 year embargo period to give graduates sufficient time to publish on their dissertation research, prior to the data becoming open access. This case supports the introduction of early data sharing incentives and guidelines for graduate students, demonstrating the different levels of open access. The embargo allows the support of Findability, Accessibility, and Reusability because the dataset metadata remains visible to the research community and the Terms of Access detail

23 For details see: https://gking.harvard.edu/software

when data will become available for public consumption.

This paper richly demonstrates how the Dataverse Software, individual installation workflows, and additional curation and data management by data depositors, can enhance the FAIR principles in Dataverse repositories. We demonstrate the vast diversity in Dataverse data sharing options that support FAIR, and provide examples of opportunities to educate researchers in the FAIR data sharing process. While the software provides functionality to move data towards FAIR, the researcher, data manager, and curator's use of the software features is what allows maximum Findability, Accessibility, Interoperability, and Reusability of the shared research content.

5. APPENDIX - Glossary (terms definitions used in this paper)

Bidirectional linking (via related publications metadata field): The dataset metadata field used to link to the related article that supports the data.
Dataset: a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.
Dataverse: Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.
Deaccessioning (tombstone page): the process of removing a published dataset for legal or valid reasons. This always results in a tombstone landing page with the basic citation metadata always accessible to the public if they use the persistent URL (Handle or DOI) provided in the citation for that dataset. Users will not be able to see any of the files or additional metadata that were previously available prior to deaccession.
Facets: metadata fields to use as facets for browsing datasets and dataverses in this dataverse.
File: A data file is a computer file which stores data to be used by a computer application or system, including input and output data (wikipedia).
Guestbook: GuestBooks allow you to collect data about who is downloading the files from your datasets.
Make Data Count: Make Data Count is a project to collect and standardize metrics on data use, especially views, downloads, and citations. Dataverse can integrate Make Data Count to collect and display usage metrics including counts of dataset views, file downloads, and dataset citations.
Metadata blocks: metadata based on standards , shipped with Dataverse (e.g. DDI for social science) and you can learn more about these standards in the Appendix section of the User Guide.
Publishing on Dataverse: When you publish a dataset, you make it available to

the public so that other users can browse or search for it using the DOI/Handle or metadata.

Tabular Data: dataverse software extracts the data content from the user's tab files and archive it in an application-neutral, easily-readable format.[24]

Versioning: Versioning is important for long-term research data management where metadata and/or files are updated over time. It is used to track any metadata or file changes (e.g., by uploading a new file, changing file metadata, adding or editing metadata) once you have published your dataset.

Widget: The Widgets feature provides you with code for your personal website so your dataset can be displayed. There are two types of Widgets for a dataset: Dataset Widget and the Dataset Citation Widget.

Templates: Templates are useful when you have several datasets that have the same information in multiple metadata fields that you would prefer not to have to keep manually typing in, or if you want to use a custom set of Terms of Use and Access for multiple datasets in a dataverse.

6. References

CROSAS, Mercè. FAIR principles and beyond: Implementation in Dataverse. European Dataverse Workshop 2020. Available at: < https://scholar.harvard.edu/files/mercecrosas/files/fairdata-dataverse-mercecrosas.pdf >. Accessed 15 Oct. 2020.

DAVID, Romain et al. FAIRness Literacy: the Achilles' Heel of applying FAIR Principles. 2020. Available at: < https://datascience.codata.org/articles/10.5334/dsj-2020-032/ >. Accessed 15 oct. 2020.

GO-FAIR. FAIR Principles. Available at: < https://www.go-fair.org/fair-principles/>. Accessed 23 set. 2020.

JACOBSEN, Annika et al. FAIR principles: interpretations and implementation considerations. 2020. Available at: < https://www.mitpressjournals.org/doi/abs/10.1162/dint_r_00024 >.

The Dataverse Project. Available at: < https://dataverse.org/about >. Accessed 30 set. 2020.

KING, Gary. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research, 36, p. 173–199.2007. Available at: < https://j.mp/2owjuRr >. Accessed 04 mar. 2021.

NAP (NATIONAL ACADEMIES OF SCIENCES). Reproducibility and

24 The supported formats are listed here: https://guides.dataverse.org/en/5.1.1/user/tabulardataingest/supportedformats.html

Replicability in Science. Available at: < https://www.nap.edu/resource/25303/R&R.pdf >. Accessed 10 oct. 2020.

ROCHA, Rafael Port da et al. Relatório de acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas. Porto Alegre, RS: UFRGS, 2018. 75 p. Available at: < https://lume.ufrgs.br/bitstream/handle/10183/185126/001082284.pdf?sequence=1&isAllowed=y >. Accessed 27 feb. 2021.

WILKINSON, Mark D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016.

Como citar com o DOI individual

REZENDE, Laura Vilela Rodrigues; BARBOSA, Sonia. Using the DATAVERSE project to move towards fair principles. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 31 -46. DOI: 10.22477/9786589167242.cap3

Rumo à rede de implantação GO FAIR 'Agro' Brasil: a experiência de uma organização de PD&I na implantação dos princípios FAIR

Debora Pignatari Drucker[1], Juliana Meireles Fortaleza[2], Patrícia Rocha Bello Bertin[3], Isaque Vacari[4] e Carla Geovana do Nascimento Macario[5]

1. Introdução

As Ciências Agrárias são multi e interdisciplinares devido à natureza integrativa dos conhecimentos sobre os meios físico, biológico, social e econômico. Além disso, a produção de alimentos, seu comércio e consumo estão intimamente ligados à saúde humana e do ambiente, sendo que a habilidade de conectar conhecimentos sobre esses temas, que muitas vezes são controversos, é essencial para a superação dos problemas complexos da agricultura (HILIMIRE, 2016). Considerados a base do método científico (HEIDORN, 2008), dados são essenciais para a análise integrada das diferentes disciplinas que compõem a pesquisa agropecuária, devendo ser tratados como produtos valiosos da atividade de PD&I.

Nesse contexto, a quantidade de dados passíveis de análise vem aumentando exponencialmente, com a transição para uma ciência intensiva em dados em todas as áreas do conhecimento (HEY *et al.*, 2009). No âmbito das Ciências Agrárias, é digno de nota o uso de sensores remotos e proximais para monitorar variáveis de interesse em tempo real, sejam de solo, clima, atmosfera ou de organismos, bem como a produção de grandes volumes de dados com novas tecnologias como as genômicas ou o processamento de linguagem natural. Ao mesmo tempo, dados com características típicas de "cauda longa", isto é, hete-

1 Doutora em Ambiente e Sociedade, Embrapa Agricultura Digital, debora.drucker@embrapa.br
2 Mestre em Fitotecnia, Secretaria de Desenvolvimento Institucional da Embrapa, juliana.fortaleza@embrapa.br
3 PhD em Gestão da Informação, Secretaria de Desenvolvimento Institucional da Embrapa, patricia.bertin@embrapa.br
4 Mestre em Ciência da Computação, Embrapa Agricultura Digital, isaque.vacari@embrapa.br
5 Doutora em Ciência da Computação, Embrapa Agricultura Digital, carla.macario@embrapa.br

rogêneos, diversos, pouco estruturados, difíceis e custosos para serem obtidos (HEIDORN, 2008; BORGMAN *et al.*, 2016), são acumulados há décadas – registros factuais valiosíssimos sobre os sistemas de produção e patrimônio ambiental, que estão sob o risco de serem perdidos caso não sejam adequadamente tratados e preservados.

Confiabilidade e reprodutibilidade são também pilares do método científico e, para assegurá-las, é fundamental que boas práticas de gestão de dados de pesquisa sejam adotadas. Nesse contexto, os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) – acrônimo em inglês para dados "localizáveis, acessíveis, interoperáveis e reutilizáveis" (WILKINSON *et al.*, 2016), vêm sendo amplamente adotados como norteadores da gestão de dados de pesquisa e viabilizadores de seu reúso. Aplicações e métricas para adoção concreta dos princípios FAIR têm sido desenvolvidas em todo o mundo, por meio da adoção de padrões, metadados, vocabulários controlados, ontologias e identificadores persistentes que proporcionam significado preciso aos dados e aos demais objetos a eles vinculados (HENNING *et al.*, 2019).

O reconhecimento pela comunidade científica da importância de tornar as práticas de gestão de dados aderentes aos princípios FAIR motivou o surgimento de iniciativas como a GO FAIR, que articula comunidades de prática a partir de um escritório central na Europa e Redes de Implementação (RIs) temáticas e regionais[6]. Uma dessas redes, intitulada *Food Systems*, tem por objetivo apoiar a implementação dos princípios FAIR nas ciências agroalimentares[7]. O Brasil integra a iniciativa com um escritório nacional[8], ao qual estão associadas diversas RIs - entre elas, a rede GO FAIR Brasil Agro, que contribuirá para a adoção dos princípios FAIR no contexto de instituições produtoras de dados agropecuários (*Go-Change*); a realização de treinamentos e capacitações , em parceria com outras RIs nacionais (*Go--Train*); e a construção colaborativa e implementação de infraestrutura e padrões intercambiáveis (*Go-Build*).

O objetivo deste capítulo é relatar a experiência da Embrapa na incorporação dos princípios FAIR às diretrizes institucionais, aos processos e às práticas de governança e gestão de dados de pesquisa. A narrativa foi construída a partir de estudo de caso com abordagem exploratória, tendo como única unidade de análise a Em-

6 O portal GO-FAIR contém mais informações: https://www.go-fair.org Mais informações em: https://www.rd-alliance.org/.

7 A Rede de Implementação Food Systems está descrita em https://www.go-fair.org/implementation-networks/overview/food-systems.

8 Mais informações sobre o escritório brasileiro podem ser encontradas em: https://www.go--fair-brasil.org/.

brapa e os dados coletados por pesquisa documental. Com fundamentação teórica nas conceitualizações da Ciência Aberta, da e-Science e da gestão de dados de pesquisa sob a ótica dos princípios FAIR, espera-se que a análise ora apresentada sirva de base para a construção da rede de implementação GO FAIR Brasil Agro, de modo a beneficiar o sistema nacional de PD&I agropecuário com as lições aprendidas no caso Embrapa.

As próximas seções apresentam informações contextuais sobre a Gestão de Dados de Pesquisa (GDP) na Embrapa, explicitam o posicionamento da Empresa no ecossistema global de GDP, descrevem os normativos internos existentes na temática e relatam os resultados alcançados até o momento. Ao final, são descritos os desafios e perspectivas futuras para estabelecer as bases para tornar os dados de pesquisa da Embrapa cada vez mais em conformidade com os princípios FAIR.

2. Um pouco de contexto: a gestão de dados de pesquisa na Embrapa

A Embrapa – instituição pública de pesquisa agropecuária vinculada ao Ministério da Agricultura, Pecuária e Abastecimento (Mapa) – tem a missão de "viabilizar soluções de pesquisa, desenvolvimento e inovação para a sustentabilidade da agricultura, em benefício da sociedade brasileira" (EMBRAPA, 2020, p. 16). Estruturada em 43 centros de pesquisa distribuídos geograficamente em todo o País e com atuação no exterior, a Empresa gera um grande volume de dados nos diversos temas estratégicos da pesquisa agropecuária. Ciente do volume, da velocidade, da variedade e do valor dos dados de pesquisa produzidos no desenvolvimento de suas atividades, a Embrapa tem mobilizado esforços para governar e gerenciar adequadamente esses ativos em todo o seu ciclo de vida, a fim de torná-los localizáveis, acessíveis, interoperáveis e reutilizáveis.

Dentre esses esforços, pode-se destacar a execução, entre 2015 e 2017, do projeto "Governança de Dados e Informação para o Conhecimento na Embrapa: Desenvolvimento de Modelo e Plano de Implantação (GovIE)", que teve por objetivo conceber, validar e propor um modelo sistêmico para a governança de dados e informação na Empresa. Como resultado do projeto, foram identificadas diversas medidas a serem tomadas para o aprimoramento da governança e gestão de dados de pesquisa na Empresa (Quadro 1), as quais encontram-se em curso de implantação.

Quadro 1. Medidas necessárias para aprimoramento da governança e gestão de dados de pesquisa na Embrapa, associadas aos pilares GO Build, GO Change e GO Train, das Redes de Implantação GO FAIR.

Medidas	Recomendações
Processuais	1. Modelar e implementar os processos corporativos de gestão de dados de pesquisa e de publicação de dados abertos. *Change* 2. Modernizar o processo de avaliação de desempenho e recompensa dos empregados de forma com o objetivo de favorecer a cultura do compartilhamento e reúso de dados. *Change* 3. Desenvolver e implementar processos, competências, ferramentas e metodologias que possibilitem a interoperabilidade semântica entre sistemas de informação. *Build* 4. Possibilitar a agregação dos dados científicos às publicações e aos projetos que os geraram. *Build* 5. Adotar e operacionalizar o uso de licenças públicas para os ativos digitais. *Build*
Normativas	6. Elaborar e publicar o Plano de Dados Abertos da Embrapa. *Change* 7. Elaborar, revisar, atualizar e implementar políticas e normas internas relativas à gestão de dados e informação gerados durante a pesquisa desenvolvida pela Empresa. *Change* 8. Incluir no Plano Diretor da Embrapa (PDE) diretrizes estratégicas e específicas relacionadas à gestão de dados e informações. *C* 9. Estabelecer e prover manutenção de um modelo corporativo de dados de pesquisa. *Build*
Cultura organizacional	10. Garantir a inserção e participação ativa da Embrapa em fóruns e *networks* nacionais e internacionais nas temáticas de gestão de dados de pesquisa. *Change* 11. Revalorizar os profissionais da Ciência da Informação da Embrapa. *Change* 12. Implantar ações de capacitação e comunicação sobre gestão de dados de pesquisa. *Train* 13. Adotar a prática de elaboração de Planos de Gestão de Dados de Pesquisa. *Change*
Ferramentas, instrumentos e tecnologia	14. Desenvolver e implantar infraestrutura tecnológica para a gestão de dados de pesquisa, por meio de plataformas consistentes e interoperáveis, para uso corporativo. *Build* 15. Adotar identificadores persistentes para dados, conjunto de dados e autores. *Build* 16. Garantir o alinhamento de planos de gestão de dados e arquiteturas da informação agropecuários com desenhos conceituais epistemologicamente sistematizados e globalmente utilizados. *Build* 17. Construir infraestrutura de dados abertos da Embrapa, interligada ao Portal Brasileiro de Dados Abertos. *Build* 18. Implantar ferramentas tecnológicas de gestão terminológica e alinhamento conceitual. *Build*
Estrutura, papéis e responsabilidades	19. Definir instância organizacional responsável pela governança de dados de pesquisa. *Change*

Fonte: Elaborado pelos autores.

Nota-se que as mudanças culturais (*Go Change*), a realização de treinamentos e capacitações (*Go Train*) e a construção e implementação de infraestrutura e padrões intercambiáveis (*Go Build*) permeiam as 19 recomendações do Quadro 1. As medidas processuais contam com recomendações relacionadas com a promoção de mudanças (*Change*: 1 e 2) e com a construção implementação de infraestrutura (*Build* : 3, 4 e 5), assim como as recomendações de medidas normativas número 6, 7 e 8 enquadram-se na categoria *Change*, enquanto a de número 9 é categorizada como *Build*. Já as medidas de Cultura Organizacional contam com uma recomendação de treinamento (*Train*, 12) e as demais são categorizadas como *Change* (10, 11 e 13). Todas as recomendações da medida Ferramentas, instrumentos e tecnologia são de construção e implementação (14, 15, 16, 17 e 18: *Build*), enquanto a medida de Estrutura, papéis e responsabilidades é categorizada como *Change*. No total, nove

recomendações foram categorizadas como de construção e implementação (*Build*), oito como de mudança cultural (*Change*) e uma como treinamento e capacitação (*Train*). Uma das recomendações categorizadas como mudança cultural, a número 10 ("Garantir a inserção e participação ativa da Embrapa em fóruns e networks nacionais e internacionais nas temáticas de gestão de dados de pesquisa"), entendida como uma medida de "Cultura organizacional", é detalhada na próxima seção.

3. Inserção da Embrapa no ecossistema global de gestão de dados de pesquisa

O fenômeno do Big Data e os novos paradigmas da e-Science e da Ciência Aberta têm promovido uma transformação no fazer científico, com reconfiguração de práticas, regras e comportamentos – sobretudo, na organização e gestão de dados de pesquisa (albagli *et al.*, 2015). Para melhor compreender e beneficiar-se dessa transformação, a Embrapa tem buscado intercambiar conhecimentos por meio da participação em iniciativas, redes, grupos e fóruns de discussão nacionais e internacionais que tratam sobre GDP. Como parte desses esforços, a Empresa coordenou, entre outubro de 2018 e julho de 2020, o Compromisso 3 do 4º Plano de Ação Nacional em Governo Aberto, conhecido como Compromisso pela Ciência Aberta, que visou "estabelecer mecanismos de governança de dados científicos para o avanço da Ciência Aberta no Brasil" (brasil, 2018a, 2018b). A execução do compromisso contou com a parceria de diversos órgãos governamentais e da sociedade civil, entre eles: Ministério da Ciência, Tecnologia e Inovações e Comunicações (MCTI), Fundação Oswaldo Cruz (Fiocruz), Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Mapa, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), Rede Nacional de Ensino e Pesquisa (RNP), Universidade de Brasília (UnB) e Open Knowledge Brasil (OKBR). Dentre as entregas do compromisso, destacam-se: a elaboração de documentos orientadores para padrões de interoperabilidade e aferição do grau de maturidade de abertura dos dados de pesquisa; o desenvolvimento e implantação de repositórios pilotos institucionais; ações de sensibilização e de capacitação sobre o tema; articulações com editores científicos e agências de fomento; formação de rede interinstitucional de interesse na temática; e o diagnóstico da Ciência Aberta no mundo e no Brasil. A agenda do Governo Aberto mostrou-se um ambiente adequado para o fortalecimento e ampliação de parcerias entre os diversos atores do sistema científico nacional, favorecendo a convergência de ações e evitando esforços duplicados.

Destaca-se ainda a participação ngrupo de trabalho (GT) para implantação de uma Rede de Repositórios de Dados Científicos do Estado de São Paulo, criado

pela Fundação de Apoio à Pesquisa do Estado de São Paulo (Fapesp), contemplando dados e informação das universidades públicas do estado[9]. A Embrapa Informática Agropecuária, uma das unidades de pesquisa da Empresa, integra o GT, focando no compartilhamento de dados e informações da Empresa, os quais podem servir para testes de escalabilidade e de integração de dados agrícolas à rede de repositórios. Além disso, a Empresa compartilha conhecimento técnico e contribui para acelerar as atividades desenvolvidas pelo GT em assuntos como testes de escalabilidade, avaliação de ferramentas ou procedimentos de curadoria de dados, entre outros.

No contexto internacional, a Embrapa integra a iniciativa GODAN (do inglês, *Global Open Data for Agriculture and Nutrition*), que tem por objetivo promover esforços globais para a disponibilização, o acesso e o reúso de dados relevantes em agricultura e nutrição[10]. Como parte da rede de mais de 1.110 parceiros, a Empresa em atuado na tradução de materiais instrucionais sobre gestão de dados abertos em agricultura para a língua portuguesa. A Empresa tem contribuído ainda no âmbito da *Research Data Alliance* (RDA[11]) ("Aliança para Dados de Pesquisa", tradução livre) - uma iniciativa global lançada em 2013 para incentivo ao compartilhamento aberto e ao reúso de dados de pesquisa. A RDA conta com mais de 11.000 membros na atualidade, envolvendo produtores, usuários e administradores de dados que colaboram para o desenvolvimento de soluções e de boas práticas em tópicos relevantes para a GDP. Os especialistas se reúnem em grupos temáticos, sendo que um deles é o *Interest Group on Agricultural Data* (IGAD), coordenado por representantes da FAO, USDA e Embrapa. Outro grupo internacional com coordenação brasileira é o *Professionalising Data Stewardship*, o qual reúne profissionais de todos os continentes para alcançar uma visão comum quanto à profissionalização da administração de dados de pesquisa. A Rede de Observação de Dados para a Terra (DataOne[12]) é mais uma iniciativa que contribui para a elaboração da estratégia e para a adoção de melhoraes práticas de gestão de dados na Embrapa, sendo um programa estabelecido para compreender melhor a vida na Terra e o ambiente que a sustenta, conduzido pela comunidade e que fornece acesso a dados em vários repositórios de membros, promovendo as melhores práticas em gestão de dados por meio de recursos e de materiais educacionais. A interface entre a ciência e a política é o foco de atuação do *Intergovernmental Science-Policy Platform on Biodi-*

9 O portal para acesso aos dados está disponível em: https://metabuscador.uspdigital.usp.br/.

10 Mais detalhes sobre a iniciativa Godan podem ser encontrados em: https://www.godan.info/

11 O portal da RDA contém informações sobre membros e grupos de trabalho: https://www.rd-alliance.org/.

12 Mais informações em: https://www.dataone.org/.

versity and Ecosystem Services (IPBES [13]), que conta com uma força tarefa para dados e conhecimento que, dentre outras ações, propôs a política de dados da plataforma e monitora sua adoção. Em conjunto, as iniciativas mencionadas contribuem para que melhores práticas de GDP sejam disseminadas e adotadas e que, assim, os dados possam ser reusados, conduzindo ao avanço da fronteira do conhecimento e subsidiando a tomada de decisão em diversas esferas. A participação da Embrapa nesses fóruns de discussão possibilita a troca de experiências, a atualização e revisão contínua de estratégias e atividades, como as que se descrevem a seguir.

4. Política de governança de dados, informação e conhecimento da Embrapa

Uma das recomendações GO *Change*, dentre as medidas normativas apresentadas no Quadro 1, é a de no. 7: "Elaborar, revisar, atualizar e implementar políticas e normas internas relativas à gestão de dados e informação gerados durante a pesquisa desenvolvida pela Empresa". Uma ação fundamental nesse sentido foi a promulgação da Política de Governança de Dados, Informação e Conhecimento da Embrapa1 (Política GDIC) que institui princípios, diretrizes, atribuições e responsabilidades para "fortalecer os mecanismos de geração, organização, tratamento, acesso, preservação, recuperação, divulgação, compartilhamento e reuso dos ativos de informação da Embrapa" (EMBRAPA, 2019, p. 10).

Orientada pelos princípios da Constituição da República Federativa, a Declaração dos Direitos Humanos e os preceitos do movimento da Ciência Aberta, a política estabeleceu 17 princípios para a gestão de dados, informação e conhecimento na Empresa, a saber: (1) Dados, informação e conhecimento como ativos corporativos; (2) Alinhamento estratégico; (3) Desenvolvimento de capacidades e competências; (4) Infraestrutura federada; (5) Análise, inteligência e inovação baseada em dados; (6) Eficiência e economicidade; (7) Conformidade e mitigação de riscos; (8) Interoperabilidade; (9) Licenciamento; (10) Preservação e memória; (11) Privacidade, proteção e confiança; (12) Segurança; (13) Qualidade e integridade; (14) Especificidade epistemológica; (15) Aprendizagem organizacional, continuidade e retenção do conhecimento; (16) Abertura e transparência; (16.1) Acesso Aberto à informação científica; (16.2) Dados Abertos; (17) Monitoramento e responsabilidade na divulgação de informações relevantes.

Apesar de os princípios FAIR não estarem diretamente enunciados nas diretrizes e princípios que compõem a Política GDIC, o documento tem por premissa que "dados e informações bem organizados, documentados, acessíveis e verificados quanto a sua exatidão e validade são mais facilmente compartilháveis e reutilizá-

13 O portal do IPBES contém detalhes sobre a plataforma: https://ipbes.net/.

veis", o que proporciona diversas vantagens à administração (EMBRAPA, 2019). Notadamente, a publicação que originalmente lançou os princípios FAIR (WILKINSON *et al.*, 2016) é uma das referências basilares da Política GDIC, de modo que estes se constituem em elementos transversais a todo o seu conteúdo. A título de exemplo, pode-se destacar a seguinte diretriz: "implantar e sustentar processos que garantam que dados e informações produzidos pela Empresa sejam confiáveis e facilmente recuperáveis, acessíveis, interoperáveis e reutilizáveis" (EMBRAPA, 2019).

Dentre os princípios da Política GDIC, o da 'Interoperabilidade' é aquele que melhor denota a necessidade de aplicação dos princípios FAIR. Para que seja contemplado na GDP, esse princípio requer o uso de ferramentas semânticas e padrões de dados e metadados amplamente estabelecidos e difundidos. Esse princípio é fortalecido por meio da diretriz prevista na perspectiva tecnológica que direciona para a inovação e o uso de tecnologias aliadas às tendências internacionais, por meio de serviços, como o compartilhamento e o reúso de dados com amplo atendimento à interoperabilidade.

Os princípios FAIR compreendem, assim, uma referência central para o desenvolvimento do Programa Corporativo de GDP, enunciado no item 8.1, da diretriz estratégica da Política GDIC: "implementar, sustentar e monitorar um Programa Corporativo de Gestão de Dados de Pesquisa e orientar o desenvolvimento de planos de gestão de dados no contexto de projetos de Pesquisa, Desenvolvimento e Inovação" (EMBRAPA, 2019, p. 13).

5. Programa Corporativo de Gestão de Dados de Pesquisa (GDP): Ações em Andamento

5.1 Diagnóstico sobre as práticas de gestão de dados

Ao longo da história da Embrapa, inúmeras práticas de gestão de dados e sistemas de informação foram criados, de acordo com as especificidades das diversas áreas temáticas de atuação dos diferentes centros de pesquisa da Empresa. Assim, uma das ações iniciais do Programa Corporativo de GDP foi realizar um levantamento para diagnosticá-las. Para tal, foi elaborado um questionário eletrônico que buscava retratar pontos importantes relacionados aos dados de pesquisa: caracterização dos dados, coleta e documentação, armazenamento, cópia de segurança, acessibilidade, compartilhamento e reuso, e repositórios de dados de pesquisa. O questionário foi respondido por 854 produtores de dados distribuídos entre 43 unidades de pesquisa com o objetivo de fundamentar e direcionar ações corporativas de aprimoramento da GDP da Empresa, de acordo com as melhores práticas e tendências internacionais de organização e publicação de dados.

5.2 Desenvolvimento e Implantação de Repositório de Dados Confiável e Identificadores Persistentes

O contexto atual requer uma estratégia abrangente para estabelecimento das bases para tornar os dados da Embrapa FAIR, considerando a importância de fortalecer as soluções já existentes na Empresa, como o Sistema de Informação de Experimentos da Embrapa (SIExp[14]) e a Infraestrutura de Dados Espaciais da Embrapa – GeoInfo (DRUCKER *et al.*, 2017), bem como a necessidade de acomodar dados para os quais soluções adequadas ainda não haviam sido implementadas. A complexidade e a multidisciplinaridade das Ciências Agrárias demandam soluções tecnológicas que possibilitem a aderência aos dados FAIR e, ao mesmo tempo, que permitam a acomodação de dados de diversas disciplinas e seus modelos de representação e padrões, com vistas a obter um conjunto de descrição central, assim como o tratamento de especificidades via pequenas extensões deste conjunto, para que se viabilize a interoperabilidade dos repositórios de dados científicos de forma geral em diversos níveis. Considerando as melhores práticas mundialmente adotadas, optou-se pela implementação de um repositório de dados de pesquisa confiável como solução para organização, tratamento, preservação e publicação dos dados produzidos pela Embrapa.

O Repositório de Dados da Embrapa, denominado Redape, foi desenvolvido e implantado a partir do software web de código aberto *Dataverse*[15]. Adicionalmente, adquiriu-se e viabilizou-se uma infraestrutura computacional própria para o armazenamento dos dados de pesquisa, e designou-se uma equipe responsável pela administração técnica do repositório Redape, hospedado no *Data Center* Científico da Embrapa, sediado na cidade de Campinas, SP, que abriga características essenciais de segurança da informação, tais como: acesso restrito aos computadores, servidores e discos de armazenamento dos dados, bem como defesa contra ataques ao repositório, prevenção do acesso de indivíduos não autorizados a dados restritos, dentre outros.

O Redape suporta o serviço de atribuição de identificadores persistentes, um dos requisitos fundamentais para tornar os produtos de dados disponíveis para a comunidade científica. Um identificador persistente (do inglês *persistent identifier* (PID)) possibilita a identificação unívoca de um conteúdo digital e destina-se a ser uma maneira permanente de identificar e de acessar esse recurso específico. O tipo de PID mais amplamente adotado no meio científico é o DOI (do inglês, *digital object identifier*), o qual gera um link persistente que aponta para o repositório

14 Mais informações em: http://www.embrapa.br/siexp.

15 Disponível em: https://dataverse.org.

ou para outra localidade digital ao incluir a URL nos metadados. Isso fornece um sistema para a identificação persistente e acionável, bem como para o intercâmbio interoperável. O Redape está sendo validado pela implantação de estudos de caso piloto representativos da pesquisa agropecuária.

5.3 Representação do Conhecimento

De acordo com Meadow *et al.* (2007), a recuperação da informação é um processo de comunicação entre os autores e criadores de registros e os leitores. Esse processo depende de um controle adequado da linguagem (código) entre emissor e receptor e entre os documentos e as requisições dos usuários (JANAITE NETO; FERNEDA, 2016). A construção dos vocabulários controlados visa às atividades de indexação, armazenamento e recuperação da informação, representando conceitos significativos de algum domínio do conhecimento e, se possível, estabelecendo relações entre tipos e até subtipos do domínio (CHANDRASEKARAN *et al.* 1999; CINTRA, 2002; JACOB, 2003; FUJITA, 2004).

De acordo com a Plataforma Lattes do CNPq, as Ciências Agrárias podem ser subdivididas nas seguintes subáreas: Agronomia, Recursos Florestais e Engenharia Florestal, Engenharia Agrícola, Zootecnia, Medicina Veterinária, Recursos Pesqueiros e Engenharia de Pesca, Ciência e Tecnologia de Alimentos (CNPQ, 2021). Essa diversidade de subáreas contribui para o elevado número de termos que podem ser utilizados para indexação, armazenamento e recuperação da informação. O Agrovoc Multilingual Thesaurus, por exemplo, contabiliza 33.388 termos principais e 2.254 termos alternativos, incluindo alimentos, nutrição, agricultura, silvicultura, pesca, nomes científicos e comuns de animais e plantas, meio ambiente, noções biológicas, técnicas de cultivo de plantas, entre outros. O Agrotermos - vocabulário controlado construído por um grupo de trabalho permanente da Embrapa - reuniu aproximadamente 245 mil termos pertinentes ao domínio do conhecimento agropecuário a partir da reunião das terminologias em língua portuguesa encontradas em tesauros agrícolas nacionais e internacionais. A expectativa é de expansão do Agrotermos para um espaço conceitual do conhecimento agropecuário brasileiro e promover uma melhor interoperabilidade entre os sistemas de informação internos e externos.

5.4 Plano de Gestão de Dados

O planejamento de GDP é uma das etapas mais importantes no processo de desenvolvimento da pesquisa, pois é nesse momento em que são discutidos como os dados serão tratados durante todo o seu ciclo de vida e como garantir que eles estejam livremente disponíveis - respeitando a privacidade - e passíveis de serem

reutilizados, sob condições e licenças específicas claramente definidas, e que possam ser devidamente citados e referenciados. O Plano de Gestão de Dados (PGD) é, portanto, uma ferramenta fundamental para que boas práticas de gestão de dados sejam aplicadas durante o desenvolvimento da pesquisa até a publicação dos dados. Cientes da importância do PGD, agências de fomento dos Estados Unidos, União Europeia, Holanda, Reino Unido, Austrália, Canadá e Finlândia têm exigido que os projetos de pesquisa venham acompanhados de um PGD alinhado com os princípios FAIR (AVENTURIER, 2017). No Brasil, a Fapesp foi a primeira agência de fomento nacional a anunciar, em 2017, a obrigatoriedade do PGD para as solicitações de financiamento de projetos de pesquisa. As instituições de pesquisa devem inserir em seu processo de desenvolvimento da pesquisa o PGD, não apenas para a obtenção de financiamento em atendimento às agências de fomento, mas também para garantir que os dados sejam devidamente gerenciados. A Embrapa está implantando ações para Gestão de Dados de Pesquisa de seus projetos. Uma delas é exigir um PGD quando da sua submissão no sistema gestor da pesquisa da empresa, tendo o proponente de responder questões como tipo de dados, repositório a ser usado e acesso e compartilhamento. A Embrapa Informática Agropecuária adota essa prática desde 2018. Corporativamente, espera-se que, a partir de 2021, essa prática seja aplicada por todas as unidades de pesquisa, em conformidade com a diretriz estratégica da PGDIC.

6. Considerações Finais

Esse trabalho descreveu esforços que vêm sendo realizados na Embrapa para implantar a gestão de dados de pesquisa fundamentada nos princípios norteadores FAIR e procurou enquadrar as medidas mapeadas de acordo com os pilares da iniciativa GO FAIR. A seção que descreveu a inserção da Embrapa no ecossistema global de GDP demonstrou que há inúmeras frentes de atuação nessa temática, que atende uma das recomendações categorizadas como mudança cultural (*Go-Change*), a número 10: "Garantir a inserção e participação ativa da Embrapa em fóruns e networks nacionais e internacionais nas temáticas de gestão de dados de pesquisa". Os resultados obtidos até o momento são notórios e têm potencial de ser multiplicados e expandidos nos próximos anos, a partir do estreitamento de relações e laços com instituições parceiras. Essa é uma característica que embasa o pilar *Go-Change*, uma vez que a construção de comunidades fundamenta a iniciativa GO FAIR.

Outra ação de extrema importância para incitar a adoção de práticas aderentes aos princípios FAIR e sustentar a mudança cultural necessária para promover os pilares do movimento GO FAIR foi a promulgação da Política de Governança de Da-

dos, Informação e Conhecimento da Embrapa, aqui descrita. Para incorporar esses princípios no dia a dia da organização, a empresa está implantando o Programa Corporativo de Governança de Dados de Pesquisa, que garantirá os meios, serviços e ferramentas necessários para que os dados produzidos pelos projetos sejam facilmente localizáveis, acessíveis, interoperáveis e reutilizáveis. Dentre as ações que estão em curso para a implantação desse programa corporativo, estão: a realização de um diagnóstico sobre as práticas de gestão de dados; a implantação de um repositório de dados confiável, com a atribuição de identificadores persistentes e que viabiliza a descoberta de dados, até então, desconectados; o desenvolvimento de ações para viabilizar a representação do conhecimento das Ciências Agrárias e o estabelecimento da prática de elaboração de planos de gestão de dados em projetos de pesquisa desenvolvidos pela empresa.

Vale destacar, como desafio e perspectiva futura, a necessidade de mapeamento e descrição formal do processo de gestão de dados de pesquisa, para que se alcance um entendimento mais aprofundado das práticas existentes, de modo que serviços e soluções a serem ofertados seja aderente à cultura organizacional e epistemológica. Os desafios, entretanto, não são elementares, tendo em vista a característica multidisciplinar das Ciências Agrárias e a necessidade de envolver diferentes atores e competências para a consecução de uma modelagem dessa complexidade. Em associação, ações de treinamento e capacitação são de extrema importância para o sucesso da implantação do Programa Corporativo de GDP.

Outro desafio a ser enfrentado é contemplar adequadamente o princípio da interoperabilidade pela adoção de padrões de dados e de metadados - requisito fundamental para permitir que os dados sejam corretamente interpretados e, assim, viabilizar seu reúso. Novamente, considerando a grande diversidade e heterogeneidade dos dados gerados e analisados no contexto das Ciências Agrárias, trata-se de um desafio que requer a participação dos diversos atores das diferentes disciplinas que compõem a pesquisa agropecuária. A atribuição de licenças que esclareçam os termos de uso dos dados é também condição fundamental.

Como demonstrado no caso da Embrapa, a aderência aos princípios FAIR é de fundamental importância para a interoperabilidade tecnológica e semântica de dados no contexto da agricultura. A estratégia aqui apresentada parte da premissa de que os dados são produtos valiosos da atividade de pesquisa, e denota a transição para uma práxis na qual o reúso de dados da pesquisa agropecuária a partir do paradigma dos princípios FAIR é encorajado. Nesse sentido, a construção da GO FAIR Brasil Agro é de fundamental importância para que o trabalho colaborativo beneficie mutuamente as comunidades de prática de gestão de dados afeitas à temática da gestão de dados. Assim, a promoção do reúso de dados de pesquisa agropecuários

contribuirá com a solução de problemas da sociedade brasileira e mundial, considerando a relevância do País no contexto dos sistemas alimentares.

Como desafios e perspectivas futuras, a modelagem de processos de gestão de dados foi iniciada na Empresa e envolve inúmeros atores de diferentes competências e áreas de atuação, trazendo complexidade para seu desenho e melhoria. Além disso, ações de treinamento e educação são de extrema importância para o sucesso da implantação do Programa Corporativo de GDP. Outro desafio a ser enfrentado é contemplar adequadamente a interoperabilidade pela adoção de padrões de dados e de metadados, requisito fundamental para permitir que os dados sejam corretamente interpretados e, assim, viabilizar seu reúso. Novamente, considerando a grande diversidade e heterogeneidade dos dados gerados e analisados no contexto das Ciências Agrárias, trata-se de um desafio que requer a participação dos diversos atores das diferentes disciplinas que compõem a pesquisa agropecuária. Além disso, a atribuição de licenças de uso para os dados é também pré-requisito para que os tipos de reúso permitidos sejam conhecidos.

Por fim, a elaboração de estratégias de monitoramento do Programa Corporativo de GDP com vistas à aderência aos princípios FAIR, permitindo a incorporação de melhorias, é uma perspectiva fundamental para assegurar seu sucesso. Uma referência de base é o trabalho realizado no âmbito do Marco 9 do Compromisso pela Ciência Aberta, intitulado "Proposição de conjunto de indicadores para aferição da maturidade em Ciência Aberta". Apesar de os princípios FAIR não necessariamente implicarem em abertura de dados, o conjunto de indicadores para aferição do grau de maturidade de abertura de dados científicos fornece critérios objetivos que podem ser utilizados também para mensuração de sucesso nos Eixos Governança, Cultura Organizacional, Gestão de Dados Científicos e Infraestrutura Tecnológica (Fortaleza *et al.* 2020). Indicadores mais específicos poderão ser desenvolvidos , tais como: métricas de catalogação de dados de pesquisa em repositórios institucionais; quantitativo de acesso aos recursos disponibilizados; descrição e implantação de processos; adoção de licenças de uso; e estabelecimento de mecanismos de recompensa ao compartilhamento e reúso de dados.

As ações aqui apresentadas sinalizam alguns passos de um caminho para promover a transição para a lógica de que dados são produtos valiosos das atividades de pesquisa e na qual o reúso de dados da pesquisa agropecuária a partir do paradigma dos princípios FAIR é encorajado. Nesse sentido, a construção da GO FAIR Brasil Agro é de fundamental importância para que o trabalho colaborativo beneficie mutuamente as comunidades de prática de gestão de dados afeitas à temática da gestão de dados. Assim, a promoção do reúso de dados de pesquisa agropecuários contribuirá com a solução de problemas da sociedade brasileira e mundial, consi-

derando a relevância do país no contexto dos sistemas alimentares.

7. Referências

ALBAGLI, Sarita; MACIEL, Maria Lucia; ABDO, Alexandre Hannud (org.) **Ciência aberta, questões abertas**. Brasília: IBICT; Rio de Janeiro: UNIRIO, 2015.

AVENTURIER, Pascal. Plano de Gestão de Dados: uma introdução. *In:* AVENTURIER, Pascal. **A publicação científica**: blog do Pascal Aventurier sobre as publicações científicas e os dados de pesquisa. Avignon, 17 maio 2017. Disponível em: https://publicient.hypotheses.org/1660. Acesso em: 30 out. 2020.

BORGMAN, Christine L. *et al.* data management in the long tail: science, software, and service. **International Journal of Digital Curation,** Edinburg, v. 1, n. 1, p. 128- 148, Oct. 2016. DOI 10.2218/ijdc.v11i1.428.

BRASIL. Controladoria Geral da União. Inovação e governo aberto na ciência - monitoramento e execução: compromisso 3. Estabelecer mecanismos de governança de dados científicos para o avanço da ciência aberta no Brasil. 2018a. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/compromisso-3-docs/inovacao-e-governo-aberto-na-ciencia-monitoramento-e-execucao. Acesso em: 3 mar. 2021.

BRASIL. Ministério da Transparência e Controladoria-Geral da União. Secretaria de Transparência e Prevenção da Corrupção. Diretoria de Transparência e Controle Social. Coordenação-Geral de Governo Aberto e Transparência. **4º Plano de Aç**ão Nacional em Governo Aberto. Brasília, DF, 2018b. Disponível em: http://governoaberto.cgu.gov.br/esta-aberta-consulta-publica-do-4o-plano-de-acao-nacional-para-governo-aberto/4o-plano-de-acao-nacional_portugues.pdf. Acesso em: 6 out. 2020.

CHANDRASEKARAN, B.; JOSEPHSON, John R.; BENJAMINS, V. Richard. What are ontologies, and why do we need them? IEEE **Intelligent Systems**, v. 14, n. 1, p. 20-26, Feb. 1999. DOI 10.1109/5254.747902.

CINTRA, Anna Maria Marques; TALAMO, Maria de Fatima Gonçalves Moreira; LARA, Marilda Ginez Lopes de; KOBASHI, Nair Yumiko (Org.). **Para entender as linguagens documentárias**. 2. ed. São Paulo: Polis, 2002.

CNPQ. **Diretório dos grupos de pesquisa no Brasil**: áreas do conhecimento –Ciências Agrárias. Brasília, DF, [2021]. Disponível em: http://lattes.cnpq.br/web/dgp/ciencias-agrarias. Acesso em: 3 mar. 2021.

DRUCKER, Debora Pignatari *et al.* GeoInfo - infraestrutura de dados espaciais abertos para a pesquisa agropecuária. RECIIS**:** Revista Eletrônica de

Comunicação, Informação & Inovação em Saúde, Rio de Janeiro, v. 11, p. 1-17, 2017. Suplemento. Disponível em: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1083246/1/GeoInfo.pdf. Acesso em: 1 mar. 2021.

EMBRAPA. Política de Governança de Dados, Informação e Conhecimento da Embrapa. **Boletim de Comunicação Administrativa**, Brasília, DF, ano 45, n. 16, p. 1-19, 5 abril. 2019. Disponível em: https://www.embrapa.br/politica-de-governanca-de-dados-informacao-e-conhecimento. Acesso em: 3 mar. 2021.

EMBRAPA. Regimento das Secretarias da Embrapa. **Boletim de Comunicação Administrativa**, [S.l.], ano 44, n. 8, p. 1-26, 1 fev. 2018. Disponível em: https://www.embrapa.br/documents/10180/1546282/Regimento+das+Secretarias+da+Embrapa/d629c401-d2e6-fd8d-5154-ccbaaa1e3313. Acesso em: 30 out. 2020.

EMBRAPA. VII **Plano Diretor da Embrapa 2020–2030**. Brasília, DF: Embrapa, 2020. 31 p. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/217274/1/VII-PDE-2020.pdf. Acesso em: Acesso em: 3 mar. 2021.

FORTALEZA, J.M.; BERTIN, P. R. B.; DRUCKER, D.P.; ASSIS, T.B.; COSTA, M.P. Conjunto de indicadores para aferição do grau de maturidade de abertura dos dados científicos. Brasília, DF: Embrapa, CNPq, OKBR, Ibict, MCTI, 2020. 14 p.

FUJITA, Mariângela Spotti Lopes. A leitura documentária na perspectiva de suas variáveis: leitor-texto-contexto. **DataGramaZero**: Revista de Ciência da Informação, Rio de Janeiro, v. 5, n. 4, ago. 2004. https://www.brapci.inf.br/index.php/article/download/7646. Acesso em: 30 out. 2020.

HEIDORN, P. Bryan. Shedding light on the dark data in the long tail of science. **Library Trends**, Baltimore, v. 57 n. 2, p. 280-299, Fall 2008. DOI 10.1353/lib.0.0036.

HENNING, Patrícia Corrêa *et al.* Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia,** Paraíba, v. 14, n. 3, p. 175-192, 2019. DOI 10.22478/ufpb.1981-0695.2019v14n3.46969.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin, (ed.). **The fourth paradigm:** data-intensive scientific discovery. Redmond: Microsoft Research, 2009. Disponível em: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf. Acesso em 1 set 2020.

HILIMIRE, Kathleen. Theory and practice of an interdisciplinary food systems curriculum. NACTA **Journal,** Rupert, v. 60, n. 2, p. 227-233, June 2016. DOI 10.2307/nactajournal.60.2.227.

JACOB, Ellin K. Ontologies and the semantic web. **Bulletin of the American Society for Information Science and Technology**, Washington, DC, v. 29, n. 4,

p. 19-22, Apr./May 2003. DOI 10.1002/bult.283.

JANAITE NETO, Jorge; FERNEDA, Edberto. Ontologia como recurso de padronização terminológica. **Informação em Pauta**, Fortaleza, v. 1, n. 1, p. 30-45, jan./jun. 2016. DOI 10.32810/2525-3468.ip.v1i1.2016.2967.

MEADOW, Charles T.; BOYCE, Bert R.; KRAFT, Donald H.; BARRY, Carol. **Text information retrieval system.** 3rd ed. Amsterdam: Elsevier, 2007. Disponível em: https://diglibrary.weebly.com/uploads/1/8/5/1/18511482/text_info_retrieval_system.pdf Acesso em: 30 out. 2020.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, London, v. 3, p. 1-9, 2016. DOI 10.1038/sdata.2016.18.

Como citar com o DOI individual

DRUCKER, Debora Pignatari; FORTALEZA, Juliana Meireles; BERTIN, Patrícia Rocha Bello; VACARI, Isaque; MACARIO, Carla Geovana do Nascimento. Rumo à rede de implantação GO FAIR 'Agro' Brasil: a experiência de uma organização de PD&I na implantação dos princípios FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 47-62. DOI: 10.22477/9786589167242.cap4

Princípios FAIR e a gestão de bases governamentais: análise do compartilhamento de dados de registros civis por meio da iniciativa GovData

Cláudio José Silva Ribeiro[1] e Ana Cristina Meirelles Velho[2]

1. Introdução

O uso de informação e conhecimento tem impulsionado projetos de investigação, tanto em ambientes acadêmicos quanto em organizações empresariais e de governo. Presenciamos o surgimento de uma economia informacional e cada vez mais globalizada (CASTELLS, 1999).

Pode-se afirmar que estamos no tempo do ciberespaço, onde a colaboração, a instantaneidade e a fluência digital crescentemente se fazem presentes no cotidiano (BARRETO, 2014). Impulsionados pela noção da "avalanche de dados", os esforços em reunir informações para apoiar a gestão de negócios (*Business Intelligence* – BI) têm se deslocado na direção de soluções de *Big Data/Analytics*, causando reflexos nas ações em desenvolvimento na atualidade.

Desde Castells, percebe-se que a economia informacional propõe o uso de dados e informação na obtenção de resultados, logo, depreende-se que não estamos diante de uma novidade em pesquisas no campo da gestão da informação. A área de Ciência da Informação tem promovido estudos que exploram esta temática, em especial em abordagens para a gestão de ativos de informação na *Web* em níveis gerenciais, táticos e operacionais (VELHO, 2007; RIBEIRO, 2008) e na direção da curadoria digital (SAYÃO; SALES, 2013).

A identificação de padrões para metadados, interoperabilidade, compartilhamento, arquivamento, acesso, reúso de coleções, bem como o processamento e

1 Doutor em Ciência da Informação. Graduado em Engenharia. É Professor Associado da Unirio onde atua no Departamento de Processos Técnicos-Documentais e no Programa de Pós-graduação em Biblioteconomia.

2 Mestre em Ciência da Informação. Graduada em Análise de Sistemas. É Gestora de Projetos de Business Intelligence da DATAPREV

descoberta inteligente de recursos por meio de ontologias e taxonomias, passaram a fazer parte das preocupações do gestor de informação (RIBEIRO, 2014; SAYÃO; SALES, 2013). Some-se a isto que a necessidade de realizar análises preditivas nos conjuntos de dados com o uso de abordagens estatísticas, além de processos de agrupamento de grandes coleções com *data mining* e simulação, promoveram o deslocamento do gestor de informação em BI para a atividade de *Analytics* (VELHO, 2007; SIEGEL, 2013), agora sendo materializada em estruturas de *data lakes e Dataponds* (INMON, 2016). Portanto, é preciso estabelecer novos paradigmas de conhecimento para conviver com estas grandes coleções de insumos intangíveis e virtuais (LEVY, 1996).

A investigação se caracteriza como uma pesquisa descritiva, exploratória de abordagem qualitativa, com pesquisa bibliográfica, documental e estudo de caso (GIL, 2002). Partindo de uma revisão de literatura sobre as iniciativas de compartilhamento de dados governamentais, este relato apresenta o objetivo, principais componentes da plataforma GovData e um recorte para possíveis usos por diferentes entes federativos e instituições públicas, inclusive de Ensino e Pesquisa.

Além desta introdução que está sendo finalizada aqui, este artigo possui mais quatro seções: a próxima, com o arcabouço teórico utilizado como referência, seguida pelo campo empírico, pelos resultados obtidos e pelas considerações finais.

2. Compartilhando dados de governo com o uso de princípios *FAIR*

Para compreender melhor o compartilhamento de dados governamentais, é necessário retroceder alguns anos para perceber marcos, pois desde o lançamento do Portal da Transparência, em 2002, e posteriormente à chegada da Lei de Acesso à Informação (LAI), em 2011, houve a cristalização dos processos de disseminação de informação de governo.

Impulsionado pelo movimento de acesso à informação catalisado pela "Carta de Serviços ao Cidadão" e pelo *Memorandum on Transparency and Open Government* do governo americano (RIBEIRO; ALMEIDA, 2011), o governo brasileiro aderiu à iniciativa *Open Government Partnership (*OGP*)*[3], amadurecendo ainda mais os processos para disseminação, compartilhamento e reuso das informações produzidas na esfera governamental (CGU, 2012).

Como parte dos planos para a implementação da parceria de Governo Aberto, foi dado início ao Portal de Dados Abertos Brasileiros, com a disponibilização de conjuntos de dados sobre as ações de governo. Fruto de projeto conjunto entre o governo

3 Iniciativa internacional para transparência governamental e combate à corrupção. Teve seu lançamento capitaneado pelo Brasil e Estados Unidos da América (OGP, 2011; CGU, 2012).

e a sociedade, representada por organismos de padronização, universidades e organizações não governamentais, o portal foi apoiado pelas ações ligadas à Infraestrutura Nacional de Dados Abertos (INDA). Esse último esforço apresentou um conjunto de padrões, tecnologias, procedimentos e mecanismos de controle necessários para atender às condições de disseminação e compartilhamento de dados e informações públicas no modelo de Dados Abertos, em conformidade com o disposto no Projeto de Interoperabilidade no Governo (projeto e-Ping) (RIBEIRO; ALMEIDA, 2011).

Todas essas iniciativas, voltadas para o uso de tecnologias visando à otimização dos processos internos do governo, formavam o conceito de "governo eletrônico". A partir de 2015, o foco do governo passa a ser "centrado no cidadão", uma tendência mundial, e é então denominado de "governo digital" (BRASIL, [2018?]) .

Dentro do conceito abrangente de governo digital, a necessidade de implementar serviços no esteio da Transformação Digital[4] abriu caminho para alguns projetos estruturantes, dos quais destacam-se o GovData e o ConectaGov. Estes objetivam, dentre outros, viabilizar a troca de informações entre as instituições públicas delimitadas pelos decretos 8.789 de 2016 e 10.046 de 2019 (BRASIL, 2016; BRASIL, 2019), portanto, a interoperabilidade entre informações sob a guarda das instituições governamentais. A iniciativa GovData se concentra na interoperabilidade via bases de dados e a ConectaGov, via troca de dados em tempo real. Dentro do recorte deste relato, a concentração dos autores será dada na avaliação das bases de dados do GovData.

2.1 Iniciativa GovData

O decreto 8.789, de 29 de junho de 2016, estabeleceu:

> Os órgãos e as entidades da administração pública federal direta e indireta e as demais entidades controladas direta ou indiretamente pela União que forem detentoras ou responsáveis pela gestão de bases de dados oficiais disponibilizarão aos órgãos e às entidades da administração pública federal direta, autárquica e fundacional interessados o acesso aos dados sob a sua gestão, nos termos deste Decreto (BRASIL, 2016, *online*).

Posteriormente, foi revogado e substituído pelo decreto 10.046, de 9 de outubro de 2019, que manteve as mesmas orientações pertinentes a este relato, quais

4 Termo utilizado no Brasil e em outros países e que abrange projetos governamentais com uso de tecnologias digitais visando aumentar a capacidade do Estado de oferecer serviços ao cidadão e ampliar o acesso e o compartilhamento de dados.

sejam, o compartilhamento de bases de dados entre as instituições e a criação de um Comitê Central de Governança[5] de Dados. Verifica-se então que a orientação para o compartilhamento de bases de dados, para além da sensibilização para uma gestão eficiente, adquiriu força de lei.

O relatório do Ministério do Planejamento, Transição do Governo 2018-2019, apresentou 15 temas estruturantes, dentre eles o Governo Digital. Neste está apoiada a apresentação do cenário que conduziu à proposição do GovData como solução para viabilizar o acesso às bases de dados pelas instituições de forma contínua, sem requerer os protocolos de permissionamento mediante convênio que seriam necessários a cada ação.

A trajetória do Brasil no provimento de soluções tecnológicas a serviço da sociedade vem ao encontro do panorama mundial, em que vários países têm se aproximado virtualmente da população via canais remotos. Alguns países europeus e os Estados Unidos mantêm estratégias de Transformação Digital há mais de uma década, segundo a ONU. Ainda neste relatório preparatório para, na época, o futuro governo que assumiria o país em 2018, a Dinamarca era o país de referência, pela amplitude de suas iniciativas, desde serviços totalmente digitais ao compartilhamento de dados e governo aberto.

Note-se que em 2020 o Brasil obteve o primeiro lugar na América Latina como provedor de serviços digitais e, nas Américas ficou atrás apenas dos Estados Unidos (BRASIL, 2020).

A iniciativa GovData teve como sua motivação suprir a necessidade dos gestores públicos de definir políticas baseados em informações suficientes, reduzindo o empirismo de tais decisões (BRASIL, [2018?]). Assim, se o Estado reúne bases de dados relevantes que podem amparar análises para subsidiar ações de seus gestores, então tratar a dispersão de tais bases e oferecê-las de forma viável via tecnologias de acesso otimiza o uso de recursos e traz mais eficiência. Esta percepção da existência de informações relevantes, ainda não disponíveis de forma adequada, e da carência de informações pelos gestores por outro lado, formam o cerne da justificativa desta iniciativa.

A interoperabilidade de informações no governo federal foi tema do IV Fórum Nacional das Transferências da União – Compartilhamento, análise e segurança (BRASIL, 2019a). Na oportunidade, os três componentes da solução GovData foram apresentados: o *data lake*, as ferramentas de acesso e a ciência de dados. O *data lake* corresponde ao lago de dados, conjunto de bases que, conforme determinação legal, devem estar disponíveis para acesso. As ferramentas de acesso (HUE, Qlik,

5 Não faz parte do objetivo deste relato debater as possíveis semelhanças e diferenças entre os termos governança de dados e curadoria de dados.

RStudio e MicroStrategy) correspondem aos recursos disponíveis para os usuários trabalharem com as bases, viabilizando desde a produção de grandes cruzamentos de dados até painéis de visualização (*dashboards*). E, como suporte metodológico para o uso destas bases, as técnicas de ciência dos dados, a fim de garantir os resultados esperados no ciclo da análise de dados.

Em relação aos atores envolvidos na gestão do GovData, à Secretaria de Governo Digital da Secretaria Especial de Desburocratização, Gestão e Governo Digital do Ministério da Economia cabe o papel de Secretaria-Executiva do Comitê Central de Governança de Dados, que coordena as atividades dos subcomitês (BRASIL, 2019b). Para questões que transcendem o âmbito do compartilhamento de dados, o Comitê Central reporta-se ao Comitê Interministerial de Governança, instituído pelo decreto nº 9.203, de 2017 (BRASIL, 2017). Dentre os papéis definidos estão aqueles relacionados à detenção dos dados (de gestor de dados, custodiante de dados, gestor de plataforma de interoperabilidade) e relativos aos interessados no uso (recebedor de dados, solicitante de dados) (BRASIL, 2019b).

Ao analisar as características do GovData, é possível inferir que os pressupostos de compartilhamento apontados pelos princípios FAIR podem ser utilizados com o intuito de tornar as suas estruturas alinhadas ao reúso de conjunto de dados proposto para a área de C&T.

2.2 Princípios *FAIR* e os requisitos para avaliação

A motivação para a proposição de princípios para compartilhamento e reúso de dados impulsionou a formulação de princípios FAIR (*Findable, Accessible, Interoperable and* Reusable). Neste relato partiu-se do pressuposto de que esses princípios já estão disseminados, pois nos últimos anos já houve diferentes estudos cobrindo a temática FAIR no contexto das universidades brasileiras, conforme observado em Henning, Ribeiro, Sales, Moreira e Santos (2018); em Moreira, Bonino, Pires, Sinderen e Henning (2019); em Ribeiro (2019) e em Monteiro e Santana (2020).

Em essência, os princípios FAIR objetivam a interoperabilidade e o reúso de dados. Isso se dá pelo atendimento aos requisitos abaixo:

- dados e seus metadados fazendo uso de identificadores persistentes e universais;
- dados e seus metadados representados por meio de uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento;
- dados e seus metadados devem possuir referências qualificadas para outros metadados e dados. Esses elementos e suas relações precisam ser descritos semanticamente;

- dados e seus metadados devem ter a sua proveniência indicada para uso e reúso, bem como os seus processos de transformação e histórico;
- dados e metadados devem estar licenciados de forma clara;
- utilizar padrões compartilhados por comunidades.

Esses requisitos foram analisados de forma minuciosa à luz das iniciativas para avaliação identificadas em Ribeiro (2019), Monteiro e Santana (2020) e Taco de Bruin (2020). O elenco de questões utilizadas na averiguação foi estruturado segundo as visões: Organizacional, do Conteúdo Digital e Tecnológica.

A visão Organizacional cobriu os aspectos ligados à infraestrutura de gestão e governança de dados, incluindo a existência de políticas específicas e perfis apropriados para a execução das atividades de gestão.

A visão do Conteúdo Digital cobriu os aspectos ligados aos dados e metadados, estratégias de identificação, descrição semântica e indexação.

A visão Tecnológica cobriu os aspectos de padronização, infraestrutura tecnológica e protocolos de comunicação, além de serviços para coleta e disponibilização.

Adicionalmente, ao tratar a adoção de princípios FAIR, GOFAIR (20[??]) apresenta o *FAIRification Process* como caminho para possibilitar instituições transformarem seus conjuntos de dados em *datasets* alinhados a esses princípios. Esse processo está organizado nas etapas:

1) Reunir e analisar os conjuntos de dados;
2) Definir e representar o modelo semântico para os conjuntos de dados;
3) Fazer os dados "linkados";
4) Verificar o licenciamento dos dados;
5) Estabelecer os metadados para os conjuntos;
6) Publicar os recursos como dados FAIR.

3. Campo empírico: recorte de *datasets* analisados

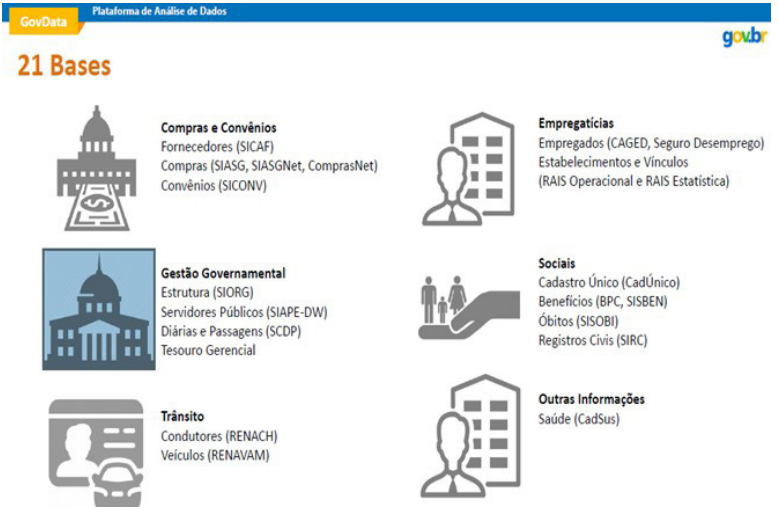
A Previdência Social Brasileira gere as informações sobre todo o ciclo de acesso do cidadão aos seus serviços, que vão desde o agendamento e orientação até a comprovação e concessão dos seus benefícios. A virtualização do acesso a estes serviços via Internet, dispositivos móveis e centrais de atendimento telefônico traz um enriquecimento ainda maior para este tema. Para cumprir a sua missão, além das informações inerentes à gestão de seus serviços, é gestora de grandes cadastros de informações sociais, que se constituem de bases de pessoas físicas e jurídicas do país, seus eventos civis (nascimento, casamento, óbito), sua

vida laborativa e eventos relacionados. Estas bases têm a sua proveniência tanto em suas próprias fontes quanto em fontes externas providas por outras instituições governamentais.

Diante da diversidade e correlações entre as suas informações, a gestão dos seus metadados é parte integrante da curadoria de suas bases de dados.

No evento já mencionado, IV Fórum Nacional das Transferências da União – Compartilhamento, análise e segurança (BRASIL, 2019a), foram apresentadas as 21 bases disponíveis na plataforma GovData à época, dentre elas, a base de dados de Registros Civis (SIRC), cujo recorte de análise será utilizado.

Figura 1 - Apresentação das bases de dados.



Fonte: BRASIL, 2019a.

O Sistema Nacional de Informações de Registro Civil (SIRC) é o meio de captação de forma digital dos dados civis de nascimento, casamento, óbito e natimortos, que são enviados pelos cartórios com o objetivo, dentre outros, de erradicar o sub-registro no país, qualificar outras bases governamentais, subsidiar políticas públicas e ajudar a coibir fraudes na concessão de benefícios e crimes como falsificação e tráfico de pessoas (BRASIL, [2019?]). Está presente no GovData com a seguinte coleção de dados:

- nascimentos: identificação do indivíduo, data e local do evento de nascimento, data e local do registro, local de residência, sexo, identificação da filiação;
- óbitos: identificação do indivíduo, nacionalidade, data e local do evento do óbito, data e local do registro, estado civil, sexo, identificação da filiação, cau-

sa da morte, data e local de nascimento, local de residência, identificador do benefício previdenciário, ocupação;
- casamentos: identificação dos cônjuges, nacionalidades, data e local do evento do casamento, data e local do registro, identificação da filiação dos cônjuges, regime do casamento, dados sobre o casamento religioso, local de residência, dados sobre a dissolução do casamento, ocupações;
- históricos: versões anteriores dos registros de nascimentos, óbitos e casamentos que tenham sofrido algum tipo de alteração;
- operacionais: codificação de dados da coleção (por exemplo, código e descrição do regime de casamento), cadastro de serventias/cartórios, dados sobre as operações de envio de arquivos para alimentar a base de dados.

Foram utilizadas para análise as coleções de dados relacionadas às certidões de nascimentos e de óbitos, pela necessidade de recorte deste relato bem como pela sua relevância como atos que representam a demografia do país.

Não foi objetivo desta investigação verificar a correlação das informações disponíveis com a legislação pertinente aos Registros Civis no país. Não obstante, reconhece-se a relevância de estudos futuros com esta finalidade.

A descrição dos conjuntos de dados disponíveis sobre os temas registro de nascimento e de óbito está disponível contendo o nome do atributo, seu formato e descrição estendida, portanto metadados técnicos. Informações sobre a sua atualidade e seu custodiante também estão disponíveis. Não há informações complementares sobre dados relacionados, não se podendo inferir se de fato não há outras relações ou se apenas há a sua omissão (CKAN, [201-]).

Em relação à atualidade do conjunto de dados, é possível verificar que a data da última atualização ocorreu há mais de um ano, possivelmente indicando a sua desativação. Entende-se, porém, que este fato não compromete a análise a que se propõe este artigo, uma vez que se trata de uma análise sobre a iniciativa implantada.

Os conjuntos de dados disponíveis sobre os temas recortados, nascimentos e óbitos, possuem grau de utilização significativo para a compreensão destes fenômenos em suas dimensões. São registros oficiais administrativos que podem apoiar estudos populacionais em conjunto com outras fontes de instituições que já são de reconhecida relevância no país, como o IBGE (IBGE, [201-]). Dados relativos ao local dos eventos podem demonstrar a movimentação da população em relação ao local de nascimento e residência. A ocupação, idade e sexo por local de residência têm juntos potencial estatístico para análises históricas de óbitos. A causa da morte, idade, sexo e local constituem um grupo de dados que combinados podem gerar indicadores de interesse para aprofundamento da realidade.

4. Resultados

Os *datasets* foram analisados à luz do apresentado na segunda seção. Os conjuntos de dados foram categorizados e reunidos para análise pelas temáticas descritas. O quadro 1 reúne as principais considerações, resultado do processo de avaliação.

Quadro 1 - Análise de atendimento a requisitos pelo GovData.

Visão	Requisito	Análise
Organizacional	Estrutura organizacional e de pessoal definida	Sim. Os *datasets* analisados são custodiados pela Dataprev. A empresa possui área específica para governança de dados.
	Papéis e responsabilidades definidas	Sim. Os *datasets* analisados são custodiados pela Dataprev. A empresa possui área específica para governança de dados.
	Políticas existentes incluem princípios *FAIR*	Não.
	Equipes dedicadas à gestão de dados, metadados e à ciência de dados.	Sim. Os *datasets* analisados são custodiados pela Dataprev. A empresa possui área específica para governança de dados e para atividades ligadas à ciência de dados.
Conteúdo Digital	Dados e metadados com identificador persistente	Parcialmente. Os identificadores dos dados e metadados não podem ser enquadrados no conceito de ID persistente.
	Metadados enriquecidos	Parcialmente. O modelo de metadados segue o estabelecido pela ferramenta CKAN.
	Dados e metadados recuperáveis pelo identificador e com protocolo padronizado, aberto e gratuito.	Sim. HTTP. Especialmente para metadados - DCAT (CKAN).
	Metadados disponíveis, mesmo após a retirada dos conjuntos de dados.	Não houve evidências.
	Dados e metadados incluem referências qualificadas para outros elementos (dados e metadados)	Não houve evidências.
	Dados e metadados estão licenciados	Metadados licenciados com Creative Commons. Não há indicação para dados.
	Dados e metadados com proveniência detalhada	Não possui.
Tecnológica	O protocolo permite procedimento de autorização e autenticação, quando necessário.	Sim.
	Dados e metadados com linguagem formal, acessível, compartilhada e aplicável para representação do conhecimento.	Metadados DCAT/RDF com protocolo Harvesting Java e compatível com OAI-PMH. Arquitetura REST. Dados com HUE/Hadoop e RStudio.
	Dados e metadados usam vocabulários que seguem os princípios *FAIR*	Não
	Dados e metadados atendem a padrões do domínio	Parcialmente. Seguem padrões locais às instituições de custódia, mas não foi percebido o alinhamento com o VCGE[6].

Fonte: elaborado pelos autores.

6 VCGE - Vocabulário Controlado de Governo Eletrônico. Disponível em: <https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico>. Acesso em: 15 out. 2020.

O resultado obtido no contexto da visão organizacional foi satisfatório, pois as instituições (Dataprev e Serpro) que fazem a gestão dos conjuntos de dados possuem alta especialização, experiência e recursos tecnológicos para a governança adequada.

Verifica-se ainda a necessidade de priorizar projetos que tratem da visão do conteúdo digital, pois os requisitos ligados à representação de metadados e o uso de modelos semânticos precisam ser atendidos em sua totalidade para permitir o reúso adequado dos conjuntos disponibilizados. Outro requisito importante é ligado ao uso de identificadores persistentes para acesso a dados e metadados. O debate sobre a geração de identificadores para dados de pesquisa pode ser estendido para abarcar também os conjuntos de dados governamentais.

De forma análoga à visão organizacional, para a visão tecnológica também foram obtidos resultados satisfatórios. Destaca-se apenas a falta de alinhamento com o VCGE. Conforme Ribeiro e Vieira (2015) apontaram, esse vocabulário foi concebido em 2011 e poderia estar em uso para incrementar a semântica dos dados.

Por fim, após cotejar os requisitos presentes no Quadro 1 com os conjuntos de dados sob análise, e tomando-se por base o *Fairification Process* apresentado na seção 2.2, verificou-se que pode ser viável a busca pelo alinhamento com princípios FAIR. Os conjuntos de dados analisados e descritos na seção 3 demonstram a riqueza de relações que podem ser construídas a partir das informações de registros civis de nascimentos e óbitos registrados nos cartórios do país.

Entende-se que os conjuntos de dados presentes requerem algum grau de anonimização para impedir a identificação dos indivíduos, mas essa transformação é possível diante da existência de insumos já disponíveis e não reduz o potencial que estas bases de dados têm de produzir novas informações.

5. Considerações finais

O compartilhamento de dados é essencial no desenvolvimento colaborativo de ações para o incremento das pesquisas. Retratando tema de interesse recente, o aumento na velocidade de obtenção de resultados ficou explícito ao ser analisado o contexto dos projetos ligados ao desenvolvimento da vacina do coronavírus (SARS-CoV-2 - COVID-19) e em especial do projeto VODAN[7].

O reúso de dados é peça-chave nesse incremento, portanto, as descrições em metadados e modelos semânticos são itens fundamentais para a compreensão adequada dos conjuntos de dados que estiverem disponibilizados.

7 *Virus Outbreak Data Network* - projeto baseado em rede FAIR para compartilhamento de dados sobre a COVID-19. Disponível em: <https://www.go-fair.org/implementation-networks/overview/vodan/>. Acesso em: 7 out. 20.

Para além do contexto da Ciência e Tecnologia, é possível inferir que os princípios FAIR também são aplicáveis no contexto do compartilhamento de dados e informações governamentais. Nesse sentido, os esforços apontados neste relato podem servir como ponto de partida para a melhor disseminação de dados e informação para a sociedade organizada e para instituições de Ensino e Pesquisa.

No momento da redação deste relato, a plataforma GovData encontra-se apresentada no site institucional do Ministério da Economia com os seus três componentes já mencionados [ver 2.1], porém sem referências às bases de dados disponíveis no *data lake*. Esse fato parece indicar aos autores que há prioridade no oferecimento dos recursos tecnológicos que propiciam o acesso e análises próprias da ciência de dados do que na disponibilização dos conjuntos de dados governamentais propriamente dita.

Por fim, a intenção deste relato é incorporar ao debate científico a possibilidade de envolvimento da sociedade nas ações de governo aberto e da construção da ciência aberta. É possível inferir que a trajetória é longa; os primeiros passos que foram dados na direção do *Citizen Analyst* (ALLEMANG, 2010; RIBEIRO; ALMEIDA, 2011) podem agora ser trilhados em busca do *Citizen Data Scientist* (BANKER, 2018).

6. Referências

ALLEMANG, D. **Nodalities:** The Magazine of Semantic Web. n. 9. 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.216&rep=rep1&type=pdf>. Acesso em: 7 out. 2020.

BANKER, S. **The Citizen Data Scientist.** Jan, 19. 2018. Disponível em: <https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist/#3c2a8dof 2702. Acesso em: 7 out. 2020.

BRASIL. **Decreto nº 8.789, de 29 de junho de 2016.** Dispõe sobre o compartilhamento de bases de dados na administração pública federal [revogado]. 2016. Disponível em: <http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2016/Decreto/D8789.htm>. Acesso em: 10 out. 2020.

BRASIL. **Decreto nº 9.203, de 22 de novembro de 2017.** Dispõe sobre a política de governança da administração pública federal direta, autárquica e fundacional. 2017. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2017/Decreto/D9203.htm>. Acesso em: 10 out. 2020.

BRASIL. **10 - Governo Digital** [2018?] Disponível em: <https://transicao.planejamento.gov.br/wp-content/uploads/2018/11/10_Governo-Digital_versão_para_publicação.pdf>. Acesso em: 2 out. 2020.

BRASIL. IV **Fórum Nacional das Transferências da União – Compartilhamento, análise e segurança.** 2019a. Disponível em: <http://

plataformamaisbrasil.gov.br/images/docs/eventos/2019/apresentacoes/Governanca_de_dados_-_SGD_ME.pdf >. Acesso em: 17 out. 2020.

BRASIL. **Decreto nº 10.046, de 9 de outubro de 2019.** Dispõe sobre a governança no compartilhamento de dados no âmbito da administração pública federal e institui o Cadastro Base do Cidadão e o Comitê Central de Governança de Dados. 2019b. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/decreto/D10046.htm>. Acesso em: 10 out. 2020.

BRASIL. SIRC – Sistema Nacional de Informações do Registro Civil. [2019?]. Disponível em: <https://sirc.gov.br>. Acesso em: 12 out. 2020.

BRASIL. **Portal gov.br.** Brasil em primeiro lugar na América Latina. 2020. Disponível em: <https://www.gov.br/pt-br/noticias/financas-impostos-e-gestao-publica/2020/07/brasil-esta-entre-os-20-paises-com-melhor-oferta-de-servicos-digitais>. Acesso em: 17 out. 2020.

CGU. CONTROLADORIA GERAL DA UNIÃO. DEFESA – Gestão da informa será o foco da Defesa no "Governo Aberto". 2012. Disponível em:
<https://www.gov.br/defesa/pt-br/assuntos/noticias/ultimas-noticias/05-12-2012-defesa-gestao-da-informacao-sera-o-foco-da-defesa-no-governo-aberto>. Acesso em: 10 out. 2020.

CKAN. **CKAN-GOVDATA**. [201-]. Disponível em: <https://ck.govdata.gov.br/>. Acesso em: 9 out. 2020.

GOFAIR. Fairification process. [201 -]. Disponível em: <https://www.go-fair.org/fair-principles/fairification-process/>. Acesso em: 15 out. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4ª ed. São Paulo: Atlas. 2002.

HENNING, P. C.; RIBEIRO, C. J. S.; SALES, L. F.; MOREIRA, J. L. R.; SANTOS, L. O. B. S. Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. Encontro Nacional de Pesquisa em Ciência da Informação. XIX ENANCIB, 2018. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/103506>. Acesso em: 5 out. 2020.

IBGE. **População.** [201-]. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao.html>. Acesso em: 10 out 2020.

INMON, B. **Data Lake architecture:** Designing the Data Lake and avoiding the garbage dump. Technics Publications. 2016.

LEVY, P. **O que é virtual?** São Paulo: Ed. 34. 1996.

MONTEIRO, E. C. S. A.; SANTANA, R. C. G. Repositórios de dados científicos na infraestrutura de pesquisa: adoção dos princípios fair. **Ciência da Informação,** v. 48, n. 3, 2019. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/136407>. Acesso em: 5 out. 2020.

MOREIRA, J. L. R.; BONINO, L.; PIRES, L. F.; SINDEREN, M. V.; HENNING,

P. Towards findable, accessible, interoperable and reusable (fair) data repositories: improving a data repository to behave as a fair data point | repositórios para dados localizáveis, acessíveis, interoperáveis e reutilizáveis (fair): adaptando um repositório de dados para se comportar como um fair data point. **Liinc em revista,** v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4817. Acesso em: 15 out. 2020.

OGP. OPEN GOVERNMENT PARTNERSHIP**.** Declaração de governo aberto. set. 2011. Disponível em: <www.opengovpartnership.org/open-government-declaration>. Acesso em: 4 out. 2020.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia (Itec),** v. 1, p. 96-105, 2014.

RIBEIRO, C. J. S. Modelo de maturidade para repositórios digitais: um caminho para sua adoção na gestão de dados de pesquisa | digital repositories maturity model: a way to its adoption in research data management. **Liinc em revista,** v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4816. Acesso em: 12 out. 2020.

RIBEIRO, C. J. S.; ALMEIDA, R. F. Dados Abertos Governamentais (Open Government Data): instrumento para exercício de cidadania pela sociedade. XII Encontro Nacional de Pesquisa em Ciência da Informação. **Anais.** 2011.

RIBEIRO, C. J. S.; PEREIRA, D. V. A publicação de dados governamentais abertos: proposta de revisão da classe sobre Previdência Social do Vocabulário Controlado do Governo Eletrônico. **Transinformação,** Campinas, v. 27, n. 1, p. 73-82, Apr. 2015 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862015000100073&lng=en&nrm=iso>. Acesso em: 12 out. 2020. http://dx.doi.org/10.1590/0103-37862015000100007.

SAYÃO, L.; SALES, L. DADOS DE PESQUISA: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Pesquisa Brasileira em Ciência da Informação,** v.8, n.2. 2013. Disponível em: <https://periodicos.ufpb.br/index.php/pbcib/article/view/18634>. Acesso em: 10 out. 2020.

SIEGEL, E. **Predictive Analytics.** Wiley. New Jersey. 2013.

TACO DE BRUIN *et al.* Do I-PASS for FAIR. A self assessment tool to measure the FAIR-ness of an organization Zenodo, Nov. 3, 2020. DOI: http://doi.org/10.5281/zenodo.4080867

VELHO, A. C. M. **A tomada de decisão na Previdência Social: uma reflexão das ações do produtor de informações da Dataprev.** Dissertação de Mestrado em Ciência da Informação – Universidade Federal Fluminense/Instituto Brasileiro de Informação em Ciência e Tecnologia. Rio de Janeiro, 2007. 137f.

Como citar com o DOI individual

RIBEIRO, Cláudio José Silva; VELHO, Ana Cristina Meirelles. Princípios FAIR e a gestão de bases governamentais: análise do compartilhamento de dados de registros civis por meio da iniciativa GovData. In: SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 63 -78. DOI: 10.22477/9786589167242. cap5

Seção 2

DADOS ACESSÍVEIS

Dados abertos da Plataforma Lattes segundo os princípios FAIR: exemplos do Extrator e Observatório de Informação da UFSC

Adilson Luiz Pinto[1], Thiago Magela Rodrigues Dias[2], Fábio Lorensi do Canto[3] e Washington Luís Ribeiro de Carvalho Segundo[4]

1. Introdução

O termo "acesso aberto" é uma tradução do inglês *Open Access* (OA), conceito relacionado ao acesso gratuito à informação científica na internet, principalmente artigos científicos revisados por pares ou publicados em uma revista científica especializada. O acesso aberto parte da premissa de que a pesquisa científica é em sua maioria financiada com recursos públicos, portanto, os seus resultados deveriam estar disponíveis e acessíveis sem custos para a sociedade. Considera que os pesquisadores não escrevem por motivação financeira, mas para maximizar a visibilidade, o uso e o impacto dos resultados de suas pesquisas. O movimento de acesso aberto defende ainda que, não obstante o processo de editoração e divulgação de um artigo envolva custos, estes devem ser incorporados aos custos gerais de pesquisa e não repassados aos leitores (SHAVELL, 2010; FREIRE, 2011).

No final da década de 1990 surgiram diversas manifestações em favor do acesso aberto. Entre as razões que impulsionaram a criação desde movimento destaca-se a crise de preços dos periódicos científicos, fenômeno que limitava ou impedia o acesso à informação científica de países e instituicões desprovidas de recursos para pagamento de assinaturas e licenças. Buscaram-se alternativas para prover acesso

1 Doutor em Documentação, Professor e Pesquisador da Universidade Federal de Santa Catarina (PGCIN/UFSC), adilson.pinto@ufsc.br

2 Doutor em Modelagem Matemática e Computacional, Professor e Pesquisador do Centro Federal de Educação Tecnológica de Minas Gerais, thiagomagela@cefetmg.br

3 Doutorando em Ciência da Informação pelo PGCIN/UFSC, fabio.lc@ufsc.br

4 Doutorando em Informática (UnB), Pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia, washingtonsegundo@ibict.br

de forma mais ampla, formando-se consórcios para aquisição de conteúdo para ser disponibilizado em portais e bases de dados (FLADUNG, 2007).

Com o desenvolvimento de ferramentas para a construção de repositórios e de bases de dados, o modelo de acesso aberto ganha consistência. Logo, diversas declarações em favor desse modelo passaram a ser publicadas, intensificando a implantação de uma infraestrutura básica de acesso aberto em níveis nacional e internacional (KURAMOTO, 2006).

O acesso aberto impulsiona o retorno dos esforços realizados em pesquisas com investimento público, tornando os resultados mais acessíveis. Independentemente dos significados que o termo encerra, o acesso aberto deve ser discutido com base em diferentes aspectos, dentre os quais se destaca o acesso à literatura ou ao conhecimento nela registrado (SUBER, 2007). É importante ressaltar que o acesso aberto ao conhecimento científico se refere tanto aos aspectos formais quanto aos informais do processo de comunicação científica (LEITE, 2016).

Nos últimos anos, as diretrizes do acesso aberto passaram a ser aplicados também aos planos de gestão de dados, tendo em vista que estes são instrumentos que orientam práticas de promoção da acessibilidade e da reutilização de dados de pesquisa. Com o intuito de possibilitar que os dados sejam mais facilmente encontráveis, acessíveis, interoperáveis e reusáveis, surgem os princípios FAIR, um acrônimo para *'Findable', 'Accessible', 'Interoperable' e 'Reusable'* (DE OLIVEIRA VEIGA et al., 2019).

Atualmente os princípios FAIR são mundialmente considerados os elementos norteadores das boas práticas de todo o processo de gestão de dados de pesquisa. Visam implementar um conjunto de metadados definidos tanto para uso por mecanismos computacionais automatizados, quanto para uso por pessoas. Estes, se forem devidamente adotados, viabilizam a interoperabilidade entre diferentes ambientes de dados (HENNING et al., 2019).

Os princípios FAIR estão presentes nas discussões e práticas contemporâneas da ciência de dados desde o início de 2014. Tiveram sua aplicação consolidada em 2017, quando a Comissão Europeia passou a exigir a adoção de planos de gestão de dados com base nesses princípios em projetos financiados por seus recursos. Desde então, os princípios passaram a ser norteadores da descoberta, do acesso, da interoperabilidade, do compartilhamento e da reutilização dos dados de pesquisa (HENNING et al., 2018).

Em De Oliveira Veiga et al. (2019) é possível encontrar uma sumarização dos princípios FAIR:

> **Localizável (**Findable), são: (a) dados e metadados precisam ter um identificador único persistente; (b) os dados devem ser descritos com metadados

ricos; (c) ter o identificador persistente para o conjunto de dados descrito nos metadados, e; (d) metadados e dados devem ser recuperáveis por meio de repositórios confiáveis;

Acessível (Accessible), são: (a) dados e metadados devem ser recuperados pelo seu identificador usando protocolos de comunicação padrão; (b) os protocolos devem ser gratuitos, abertos e suportar autenticação e autorização, e; (c) os metadados devem estar acessíveis mesmo quando os dados não estiverem mais disponíveis.

Interoperável (Interoperable), são: (a) dados e metadados devem ser codificados usando padrões de representação acordados, e; (b) dados e metadados devem usar vocabulários alinhados aos princípios FAIR e incluir referências relevantes.

Reutilizável (Reusable), são: (a) dados e metadados precisam estar associados a atributos relevantes; (b) dados e metadados devem ser liberados com licenças de uso claramente definidas; (c) metadados e dados devem estar associados à sua proveniência de forma detalhada, e; (d) dados e metadados devem atender aos padrões da comunidade.

A partir da consolidação desses quatro princípios no contexto de dados abertos e de ciência aberta, apresenta-se uma das formas pelas quais a Universidade Federal de Santa Catarina (UFSC) tem utilizado os dados disponíveis da Plataforma Lattes em seus sistemas internos de gestão.

2. Metodologia

A fonte de dados definida para este trabalho foi a base de currículos dos docentes da UFSC cadastrados na Plataforma Lattes (2.581 servidores docentes permanentes da UFSC em maio de 2020). A escolha dos currículos Lattes está relacionada ao fato de que possuem uma vasta quantidade de informações. É um recurso que permite a integração de dados científicos, profissionais e acadêmicos e assim como a atualização dos dados pode ser pelos próprios pesquisadores. Entre as principais informações contidas nos currículos destacam-se as sobre formação acadêmica, áreas de pesquisa, atuação profissional e orientações acadêmicas, além de produções técnicas e científicas.

Os currículos Lattes se tornaram um padrão nacional utilizado para a avaliação individual das atividades científicas e acadêmicas. Agregam dados de pesquisadores de todas as áreas do conhecimento, tornando a plataforma uma fonte relevante para análise e compreensão do comportamento de grupos de pesquisa (DIGIAMPIETRI et al., 2012).

Apesar de os dados dos currículos serem disponibilizados livremente, eles só podem ser visualizados individualmente por meio de uma interface de consulta disponibilizada pelo CNPq. Entretanto, essa interface é limitada e não permite sem possibilidade de agrupamentos, análises e comparações com outros currículos. Diante disso, técnicas e ferramentas para a extração de dados se fazem necessárias para análise de amplos conjuntos de dados curriculares.

Para a extração e tratamento inicial dos dados foi utilizado um *framework* denominado *LattesDataXplorer* (Dias, 2016). É uma ferramenta desenvolvida com o propósito de realizar a coleta e tratamento dos dados curriculares da Plataforma Lattes, com baixo custo computacional.

O *LattesDataXplorer* é responsável por englobar todo o conjunto de técnicas e métodos de coleta, tratamento e análise dos dados utilizados neste estudo. A extração é realizada por um componente que faz a busca e a recuperação dos currículos de cada docente a partir do identificador único de seu currículo na Plataforma Lattes. Consequentemente, com todos os currículos armazenados localmente em formato XML, a instituição maneja seus dados no Observatório de informação da UFSC (https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/) o que possibilita a gestão de dados para a Pró-Reitoria de Pesquisa e Pós-Graduação da instituição.

Figura 1: Modelo do sistema utilizado pela UFSC para extração e tratamento dos dados provenientes da Plataforma Lattes

Fonte: Dados da pesquisa, 2020.

Utilizando o *LattesDataXplorer* é possível agrupar um conjunto de currículos com base em parâmetros predefinidos. No processo de seleção dos currículos baseado em parâmetros, independentemente ou não da seção em que ele(s) seja(m) encontrado(s), os currículos são selecionados e formam um grupo para análise. Os dados são organizados em uma lista de currículos selecionados, o que não seria possível sem a utilização da estratégia adotada.

Posteriormente, tendo como entrada uma determinada listagem de currículos gerada por consultas específicas ou mesmo utilizando uma listagem global com todo o repositório local de currículos, é possível processar os dados com rotinas computacionais específicas para cada tipo de análise. Este processamento visa extrair dos currículos informações de interesse e agrupá-las em arquivos de dados pré-processados. Tal estratégia objetiva gerar conjuntos de dados específicos para a aplicação de métricas de análise, não sendo mais necessário acessar todo o conjunto de currículos em cada nova análise. Com isso, os currículos, que possuem uma grande quantidade de dados a ser processada, são acessados e tratados uma única vez. Como exemplos de arquivos de dados pré-processados, cita-se os que agrupam informações sobre orientações, formações acadêmicas, colaboração e produção científica, bem como sobre projetos de pesquisa e produção técnica. Todos os dados são disponibilizados de forma tabular.

3. Resultados

Os principais resultados são visíveis no Observatório da Informação da UFSC – https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/[5] –, um ambiente que apresenta dados históricos da instituição, sendo: (a) indicadores que são cruzados com dados do CNPq, como bolsistas de Produtividade em Pesquisa, Bolsistas de Desenvolvimento Tecnológico, Grupos de Pesquisa, Bolsistas de Apoio Técnico Pesquisa e Bolsistas de Apoio Técnico Extensão; (b) indicadores gerais da UFSC (2012-2018), como Distribuição da produtividade científica/técnica/artística e orientação, Produção científica por tipologia, Produção técnica por tipologia, Produção Artística e Orientações; (c) indicadores Tecnológicos (2000-2018), como Patentes em geral, Patentes temáticas, Patentes por grandes áreas, Patentes por inventor, Patentes por país de depósito, Patentes colaboradas por áreas, Patentes por colaboração por instituições e Patentes por áreas; (d) indicadores por departamentos de forma per capita entre 2012-2018; (e) indicadores de visibilidade na base de dados Web of Science (2012-2018), tendo como foco as Tipologias de publicação, Áreas de publicação, Instituições parceiras; Principais Financiadores, Países parceiros e Autores de maior visibilidade.

Entretanto, o propósito deste estudo é a viabilidade de todo este conjunto de informações em acesso aberto a partir dos princípios FAIR, no qual pretende-se

5 O Observatório é gerenciado pela SETIC/UFSC e por estar em constante manutenção na pandemia do COVID-19 é sugerido que para acessar as planilhas se utilize do botão direito do mouse para abrir o link em uma nova guia. Desta forma soluciona qualquer problema de manutenção do sistema e a possibilidade de ter acesso aos conteúdos gerados.

determinar de que forma o conjunto de dados do Extrator e do Observatório de informação da UFSC estão representados.

Atendendo o princípio Localizável (Findable), foi possível a identificação de um contexto de dados únicos, o Identificador único dos Currículos (http://lattes.cnpq.br/4767432940301118). Outra aspecto a ser destacado é que estes dados possuem diversos recursos e características, tais como sistemas de localização, níveis de formação e funções laborais, níveis de tipologias de produções, sejam científica, técnica ou artística, bem como tutorias em todos os níveis de formação. O identificador permanente do conjunto de dados pode ser recuperado a qualquer momento, visto que o conjunto de caracteres são únicos e constantemente vistoriados, no qual nenhum pesquisador pode ter dois meios de entrada. O ponto chave é que o sistema é atualizado por este ponto de entrada de dados todos os finais de semana.

Para o quesito Acessível (Accessible), o sistema que a Universidade Federal de Santa Catarina desenvolveu a partir do Extrator e do Observatório da informação é que mesmo que os dados sejam recuperados, deve-se manter um padrão de identificação do protocolo nos Currículos Lattes (4767432940301118), que também é de acesso aberto. O acesso à indexação dos dados é exclusivo dos pesquisadores pelo seu próprio currículo, entretanto a extração dos dados é livre, em especial por se tratar de um registro governamental, mantido pelo Ministério de Ciência e Tecnologia. Por último, os dados ficam disponíveis mesmo que os pesquisadores não os atualizem, inclusive em caso de óbito.

Em se tratando da Interoperabilidade (Interoperable), os dados trabalhados são um conjunto de metadados que cada pesquisador agrega a seu currículo. Alguns campos são passíveis de se utilizar filtros. Além disso, por serem formatados a partir do padrão XML, a interoperabilidade com outros conjuntos de dados está facilitada. Diversos outros sistemas foram gerados para tratamentos similares no Brasil, como o *Scipt Lattes* (MENA-CHALCO; CESAR JUNIOR, 2009).

Para a questão de Reutilização (Reusable) dos dados, na UFSC existe um sistema para a captação dos dados em formato XML, que vai para um extrator sistemático. Esse sistema ordena a informação segundo os dados funcionais de cada docente, não se preocupando com solapamento entre departamentos em um primeiro momento. Em um segundo momento, é possível realizar o solapamento institucional. A licença de uso dos dados é do governo, então como são dados de necessidade nacional requer que seja atualizado para possíveis avaliações (Bolsa de Produtividade, Projetos de editais, dentre outros).

4. Função dos dados para o controle da gestão institucional

A Universidade Federal de Santa Catarina, como instituição de ensino, pesquisa e extensão de nível superior precisa produzir indicadores de seu desenvolvimento científico, técnico, social e de internacionalização, bem como acompanhar a evolução histórica desses índices. Por este motivo que tem investido nos projetos desta natureza, seja para extração e até mesmo para a sua aplicabilidade em novos serviços, pesquisas e produtos nos quais vale a apena investir.

Pensando neste processo, as ações realizadas estão voltadas para a geração de competências e para a identificação de especialistas, sendo que o recurso apresentado neste trabalho serve de base para encontrar os talentos da instituição por meio dos currículos cadastrados na Plataforma Lattes e em outras plataformas abertas.

Para as Pró-Reitorias de Pós-Graduação e de Pesquisa, esse repertório de dados serve para identificar os especialistas da instituição e até mesmo os colaboradores de determinadas temáticas de estudo. Como exemplo de aplicação prática desde recurso citá-se o caso da pandemia de COVID-19 sendo possível verificar os pesquisadores internos e seus principais colaboradores no desenvolvimento de estudos de saúde pública, sanitárias e até mesmo de ciência de dados. Existem outros níveis que também podem ser explorados, como na questão de orientações em nível de graduação, mestrado e doutorado.

Há quatro níveis de conteúdo que podem ser explorados para condensar em um contexto de identificação dos talentos pelo Currículo Lattes, quais sejam: (i) os especialistas em produtividade científica; (ii) os especialistas em orientações e participações de bancas de teses/dissertações em temáticas/assuntos; (iii) os líderes dos grupos de pesquisa, e; (iv) os especialistas em produção tecnológica.

A identificação deste cenário pode ser visto de um quadro geral, combinando todas as possíveis identificações, supracitadas no parágrafo anterior, dando margem de porcentagem para cada um. Por exemplo, 25% para cada item ou estudando quais são mais importantes e dividindo as porcentagens conforme a relevância de cada item estudado.

Entretanto, independente da ordem ou item de maior relevância, pode-se identificar particularidades em cada um dos itens deste possível banco de talentos. Isso porque são dados localizáveis, acessíveis, que têm uma interoperabilidade de uso padrão, que podem ser reutilizados, também para determinar os padrões dos departamentos e centros de pesquisa, como em um sistema de geração de indicadores de ranqueamento e per capita.

Em se tratando da verificação a partir das orientações e participações de bancas de tese e dissertação, pode ser baseado em números absolutos de orientação, de participações de bancas ou ainda uma relação entre ambas.

O processo para chegar a esta particularidade de análise tem um *plus* de poder identificar as famílias científicas, como identificar os docentes que conseguem levar seus orientandos desde o nível de graduação, passando pelo mestrado e até chegar no doutorado. Isso também pode ser aplicado aos especialistas que conseguem levar seus orientandos em níveis acadêmicos mais elevados, verificando, inclusive, se os orientandos tiveram bolsas nestes períodos (COSTA; PINTO, 2016).

Seguindo com a linha de raciocínio temos os conteúdos abertos do Diretório de Grupos de Pesquisa do CNPq, que detém um rico acervo de dados sobre o desenvolvimento de grupos de pesquisa, seus participantes e colaboradores e por fim o item mais valioso, os líderes dos grupos de pesquisa. Estas informações podem ser úteis para a identificação de especialistas em áreas, temáticas e assuntos.

O uso destes dados podem ser isolados, especificamente identificando o perfil destes pesquisadores, como pode ser fundido ao conteúdo dos respectivos currículos dos pesquisadores.

O resultado desta fusão seriam os outros itens em análise, que tratam exclusivamente da dinâmica produtiva dos pesquisadores em questões científicas (artigos de periódicos, trabalhos apresentados em atas de eventos, livros publicações/editados e capítulos de livros), bem como a questão tecnológica (patentes, modelos tecnológicos, designer gráficos, dentre outros).

Pode ser vislumbrado ainda o capital científico identificando os autores mais citados dentro de suas respectivas áreas, dando mais um norte do contexto de especialidade para a área, temática ou assunto. O índice de citação pode ser gerado por meio das bases de dados disponíveis no Portal de Periódicos da Capes (livre para os sistemas federais) ou pelo Google Acadêmico.

Por último, para a identificação destes especialistas e para a construção deste banco de talentos, pode-se verificar o desempenho dos pesquisadores em publicações de acesso aberto, bem como em conteúdos comerciais, como revistas indexadas em bases de dados. O foco deste meio de averiguação de dados se dá para também identificar se os estudos de um conjunto de pesquisadores têm aderência em nível de internacionalização.

5. Conclusão

Este tipo de serviço é de fundamental necessidade para a UFSC, assim como para qualquer outra instituição de ensino, e pode-se resumir sua importância em seis pontos básicos:

1) É utilizado como suporte aos Programas de Pós-Graduação para se ter conhecimento, em tempo real, da evolução dos dados científicos, técnicos, ar-

tísticos e as orientações de seus docentes. Vale salientar que o Observatório de Informação também fornece serviços direcionados à monitoração da produção de um dado Programa de Pós-Graduação na perspectiva das colaborações com os demais dos programas de mesmo seguimento no Brasil, como já explanado no artigo "A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil" que monitorou os principais Programas de Pós-Graduação em Engenharia de Produção no período de 2008 a 2017 (DUTRA et al, 2019);

2) Explora e interopera com as aplicações FAIR dentro da instituição, acessando dados valiosos para seu planejamento, dos programas de pós-graduação às Pró-Reitorias, como uma alternativa de agregar as informações e indicadores de C&T em um só portal de busca;
3) Assim como nos demais *Scripts* de extração da Plataforma Lattes, é acessível, de fácil manuseio e interoperável. A partir dos currículos no formato XML, pode ter reuso de seus dados, como visto no Extrator e no Observatório da UFSC;
4) Serve de base informacional à Pró-Reitoria de Pesquisa, e ao projeto de criação de um laboratório de monitoramento em C&T na UFSC, dentro do Observatório da Informação,
5) Os dados gerados, tanto pelo Extrator, como pelo Observatório da UFSC podem servir como referência a outras instituições públicas, no que se refere ao monitoramento per capita de produtividade e de visibilidade de seus pesquisadores, visto que a UFSC é a melhor ranqueada em citação per capita do país, quando se leva em conta a eliminação de autocitações, e;
6) Por fim, reforça-se que a organização dos dados coletados da Plataforma Lattes à observância dos princípios FAIR serve de modelo à construção de outros sistemas e serviços que permitam a fácil recuperação de informações, assim como a projeção de novos indicadores sobre C&T de uma instituição, seus níveis de colaboração nacional e internacional, e a disponibilização para coleta automática destes dados brutos e agregados de forma aberta; sempre que não houver restrição legal.

6. Referências

COSTA, Airton; PINTO, Adilson Luiz. **De bolsista a cientista: a experiência da UFSC com o Programa de Iniciação Científica no processo de formação de pesquisadores (**1990 **a** 2012**)**. 1. ed. Florianópolis: EdUFSC, 2016. 165p .

DIAS, Thiago Magela Rodrigues. Um Estudo Sobre a Produção Científica

Brasileira a partir de dados da Plataforma Lattes. 2016. 181 f. Tese (Doutorado) - Curso de Programa de Pós-graduação, Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

DIGIAMPIETRI, Luciano Antônio et al. Minerando e caracterizando dados de currículos lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2, 2012, Curitiba. **Anais...** Curitiba: Brasnam, 2012.

DUTRA, Silvana Toriani et al. A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil. **Informação & Sociedade**, v. 29, n. 1, p. 117-136, 2019. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/44852. Acesso em: 27 de setembro 2020.

FLADUNG, Rainer. *Scientific communication*: economic analysis of the eletronic journal market. Stuttgart: Ibidem-Verlag, 2007.

FREIRE, José Donizetti. CNPq e o acesso aberto à informação científica. 2011. 275 f. Tese (Doutorado) - Curso de Programa de Pós-graduação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2011.

HENNING, Patrícia Corrêa et al. Desmistificando os Princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos Dados FAIR. In: XIX Encontro Nacional de Pesquisa em Ciência da Informação, 19, 2018, Londrina. **Anais...** Londrina: UEL, 2018.

HENNING, Patrícia Corrêa et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019.

KURAMOTO, Hélio. Informação científica: proposta de um novo modelo para o Brasil. **Ciência da Informação**, v. 35, n. 2, p.91-102, maio 2006.

LEITE, Fernando César Lima. **Gestão do conhecimento científico no contexto acadêmico: proposta de um modelo conceitual**. 2016. 240 f. Dissertação (Mestrado) - Curso de Mestrado em Ciência da Informação, Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2016.

MENA-CHALCO, Jesús Pascoal; CESAR JUNIOR, Roberto Marcondes. ScriptLattes: An open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31--39, 2009.

SHAVELL, Steven. Should copyright of academic works be abolished?. **Journal of Legal Analysis**, n. 1, v. 2, p. 301-358, 2010. Disponível em: http://jla.oxfordjournals.org/content/2/1/301.short. Acesso em: 01 out. 2020.

SUBER, Peter. **Open Access Overview: focusing on open access to peer-reviewed research articles and their preprints**. Creative Commons, 2007. Disponível em: http://legacy.earlham.edu/~peters/fos/overview.htm. Acesso em: 9 de ago. 2020.

DE OLIVEIRA VEIGA, Viviane Santos et al. Plano de gestão de dados fair: uma proposta para a Fiocruz. **Liinc em Revista**, v. 15, n. 2, 2019.

Como citar com o DOI individual

PINTO, Adilson Luiz. Dados abertos da Plataforma Lattes segundo os princípios FAIR: exemplos do Extrator e Observatório de Informação da UFSC. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 79 -90. DOI: 10.22477/9786589167242.cap6

Análise dos conjuntos de dados disponíveis no repositório COVID-19 Data Sharing/BR à luz dos princípios FAIR

Anderson Rafael Castro Simões[1], Renata Lemos dos Anjos[2], Guilherme Ataíde Dias[3]

1. Introdução

O constante uso das Tecnologias Digitais de Informação e Comunicação (TDICs) pelos mais diversos indivíduos e setores da sociedade, inclusive o acadêmico-científico, contribui para uma crescente e contínua produção de dados. O cenário acadêmico-científico configura-se tanto como um grande produtor quanto como consumidor destes, que passam a ser vistos como fontes primárias para novas investigações científicas, proporcionando o desenvolvimento da ciência.

Nessa perspectiva, Sales *et al*. (2020) afirmam que o avanço da ciência, nas mais variadas áreas do conhecimento, está fortemente vinculado à reutilização dos dados científicos, o que aponta uma demanda em gerenciá-los e preservá-los através de atividades de curadoria digital pelo tempo que for necessário, de modo a possibilitar o seu efetivo reúso em futuras pesquisas.

Esses dados, além de ganharem valor e importância em cenários sociais, políticos e econômicos, tornam-se componentes cruciais no enfrentamento de graves desafios sociais e ambientais do século xxi, mediante o seu compartilhamento (leonelli, 2019).

Neste ínterim, e considerando que em 11 de março de 2020, a Organização Mundial da Saúde (oms) declarou uma pandemia de covid-19 causada pelo coronavírus

1 Mestre em Gestão de Organizações Aprendentes pelo Programa de Mestrado Profissional em Gestão de Organizações Aprendentes da Universidade Federal da Paraíba – mpgoa/ufpb. E-mail: anderson.simoes@estudantes.ufpb.br.

2 Mestre em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba – ppgci/ufpb. E-mail: renata.anjos@academico.ufpb.br.

3 Doutor em Ciências da Comunicação (Ciência da Informação) pela Escola de Comunicação e Artes da Universidade de São Paulo - eca/usp. Professor do Departamento de Ciência da Informação da Universidade Federal da Paraíba. E-mail: guilhermeataide@ccsa.ufpb.br.

SARS-CoV-2, questões acerca do compartilhamento de dados científicos e a colaboração entre diferentes fontes de investigação vêm ganhando destaque e sendo evidenciadas. A pandemia apresenta uma necessidade real e urgente de união e esforço em escala mundial, sobretudo no âmbito científico, para que as lacunas sobre o novo coronavírus sejam solucionadas rápida e efetivamente (ALMEIDA *et al.*, 2020).

Assim, considerando esta realidade, a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), em cooperação com outras instituições, anunciou a criação do primeiro Repositório de Dados de Pesquisa (RDP) anonimizados abertos do país relacionados à COVID-19, o COVID-19 **Data Sharing/BR**[4], que, em síntese, objetiva disponibilizar os dados de pesquisa relacionados à COVID-19 no Brasil, de forma a contribuir com esta temática (FAPESP, *online*).

Nessa conjuntura, o COVID Data Sharing/BR concretiza a concepção de que o compartilhamento, uso e reúso dos conjuntos de dados disponíveis em um repositório auxiliam efetivamente a resolução dos desafios postos pela pandemia, o que evidencia o protagonismo atual dos dados científicos.

Contribuindo para que os dados sejam efetivamente reutilizados por outros pesquisadores e investigações científicas, algumas iniciativas foram propostas, destacando-se os Princípios FAIR. Estes princípios destacam-se como sendo uma abordagem cujo objetivo é fazer com que os dados sejam mais facilmente encontráveis (***Findable***), acessíveis (***A**ccessible*), interoperáveis (***I**nteroperable*) e reutilizáveis (***R**eusable*). Tais princípios incentivam, dentre outras práticas, o uso de metadados que, quando devidamente empregados, podem contribuir para aumentar a encontrabilidade, acesso, interoperabilidade e o reúso dos diferentes conjuntos de dados (DIAS; ANJOS; RODRIGUES, 2019).

Considerando a importância da colaboração entre cientistas neste momento de pandemia, e que a aderência aos Princípios FAIR pode favorecer para ampliar o compartilhamento de conjuntos de dados científicos disponibilizados em repositórios digitais, maximizando desta forma, o uso e reúso destes conjuntos, chegou-se à seguinte questão de pesquisa: **de que forma se configuram os conjuntos de dados disponíveis no repositório de dados COVID-19 Data Sharing/BR à luz dos Princípios FAIR?**

De maneira a responder à questão de pesquisa, elaborou-se o seguinte objetivo geral: avaliar a aderência dos conjuntos de dados disponíveis no repositório de dados COVID-19 Data Sharing/BR aos Princípios FAIR.

4 Disponível em: https://repositoriofapespcienciasaude.uspdigital.usp.br/. Acesso em 25 de fevereiro de 2021.

2. Ampliando o acesso a dados científicos no contexto da pandemia covid-19

O repositório de dados COVID-19 Data Sharing/BR foi fundamentado na ideia e na importância do compartilhamento e colaboração. Criado recentemente pela FAPESP em cooperação com a Universidade de São Paulo (USP), e a participação do Instituto Fleury, além dos hospitais Sírio-Libanês e Israelita Albert Einstein, localizados no estado de São Paulo. Este repositório integra inicialmente dados demográficos de 75 mil pacientes, 1,6 milhão de exames clínicos e laboratoriais e 6.500 dados de desfecho de pacientes que subsidiam pesquisas científicas sobre COVID-19. Os conjuntos de dados, disponibilizados no repositório, apresentam três categorias de informações: dados demográficos (sexo, ano de nascimento, região); dados de exames clínicos e/ou laboratoriais, e dados sobre a movimentação do paciente (internações, recuperação e óbito) (FAPESP, *online*).

Neste sentido, a iniciativa dos Princípios FAIR foi criada com a intenção de que possam servir como diretrizes para aqueles – cenário acadêmico-científico, indústria, agências de financiamento e editoras – que almejam melhorar a infraestrutura de suporte para incrementar o uso e reúso dos seus dados. Ao contrário de outras iniciativas que focam o usuário/humano, os princípios FAIR buscam aprimorar a capacidade das máquinas em encontrar e usar os dados de forma automática dando suporte à reutilização pelos usuários (WILKINSON *et al*., 2016). Os Princípios FAIR visam descrever considerações distintas para ambientes contemporâneos de publicação de dados, com relação aos suportes de: depósito, exploração, compartilhamento e reutilização manual e automatizada dos dados. (WILKINSON *et al*., 2016).

O primeiro princípio FAIR (***F****indable*) aborda a necessidade de tornar os dados localizáveis. Este é um pré-requisito fundamental para a efetivação dos outros três princípios FAIR. Um conjunto de dados deve ser identificável unicamente e de maneira persistente, permitindo a sua descoberta a qualquer tempo. Os dados devem ser descritos com metadados ricos de forma que o pesquisador possa encontrar os dados desejados, independentemente de ter acesso ao seu identificador (DIAS; ANJOS; RODRIGUES, 2019).

O princípio ***A****ccessible* aborda a necessidade de tornar os dados e metadados mais acessíveis a partir do momento em que são encontrados. Estas entidades devem estar sempre acessíveis para os usuários e/ou máquinas, para tal, é importante o uso de protocolos abertos, livres e universalmente implementáveis (GO FAIR, *online*). É recomendado que os metadados estejam disponíveis e acessíveis, mesmo quando a licença do conjunto de dados não permite o livre acesso a seu conteúdo (WILKINSON *et al*., 2016; GO FAIR, 2020, *online*).

O terceiro princípio FAIR (***I****nteroperable)*, aborda a necessidade de tornar os dados e outros ativos digitais mais interoperáveis. Essa questão está relacionada a

necessidade de integrar dados a outros conjuntos de dados e com as mais variadas aplicações ao longo do ciclo de vida. Para que seja possível essa interoperabilidade entre conjunto de dados, é importante que existam instrumentos para padronizar semanticamente os sistemas envolvidos nesse processo. Exemplos incluem tesauros e ontologias (GO FAIR, *online)*.

O quarto, e último princípio FAIR (***R**eusable)*, aborda a necessidade de tornar os dados reutilizáveis. A implementação da reutilização exige uma abordagem múltipla e permite que os dados sejam reaproveitados por novas comunidades de usuários, para novas necessidades e aplicações. Os dados, nesse sentido, podem se tornar mais valiosos para indivíduos nas mais diversas organizações, sejam comunidades de código aberto ou organizações privadas (WISE *et al.*, 2019). Recomenda-se que as políticas de acesso aos dados e metadados estejam explícitas, garantindo assim a compreensão sobre os direitos de acesso, uso e reúso, assim como os detalhes que indiquem a proveniência destes objetos (GO FAIR, *online)*.

Nesta perspectiva, ressalta-se a relevância da efetivação dos Princípios FAIR, que, quando implementados, podem resultar em inúmeros desenvolvimentos, incluindo a possibilidade de automação de processos por meio da capacidade de leitura automatizada por máquina de dados e metadados, contribuindo para ampliar sua reutilização e escalabilidade, além de proporcionar uma gestão mais rigorosa de dados e metadados com potencial para beneficiar toda a comunidade acadêmica. Os Princípios FAIR, desta forma, tornam-se uma premissa de apoio à descoberta e inovação da ciência (WILKINSON *et al.*, 2016; WISE *et al.*, 2019).

3. Percurso metodológico

A pesquisa busca avaliar a aderência dos conjuntos de dados disponíveis no repositório de dados COVID-19 Data Sharing/BR aos Princípios FAIR. Do ponto de vista do objetivo, caracteriza-se como uma pesquisa exploratória e descritiva. Quanto à abordagem do problema, possui uma abordagem estruturada em que todas as etapas do processo de investigação foram previamente determinadas. Quanto ao modo de investigação possui uma análise mista, cuja análise qualitativa foi utilizada na análise da aderência dos conjuntos de dados aos Princípios FAIR, e quantitativa no que tange à pontuação FAIRness de aderência aos princípios (RICHARDSON, 2017).

O *corpus* desta pesquisa foi constituído pelos conjuntos de dados disponíveis no repositório COVID-19 Data Sharing/BR, totalizando três conjuntos, cada um deles provenientes de cada uma das instituições colaboradoras, a saber: Grupo Fleury, Hospital Sírio-Libanês e Hospital Israelita Albert Einstein.

Para a verificação da pontuação FAIRness, utilizou-se a ferramenta *online Self-Assessment Tool to Improve the FAIRness of Your Dataset* – SATIFYD[5] proposta pela *Data Archiving and Networked Services* – DANS, que serve como uma recomendação do repositório EASY[6] de autoavaliação dos conjuntos de dados, para verificação do FAIRness antes da publicação no mesmo.

A ferramenta SATIFYD é composta por 12 questões que abordam os Princípios FAIR, igualmente divididas em seções, que são as próprias letras correspondentes ao acrônimo FAIR. Ou seja, *Findable* contempla as perguntas um, dois e três; *Accessible* contempla quatro, cinco e seis; *Interoperable* contempla sete, oito e nove e; *Reusable* contempla as perguntas dez, onze e doze.

Como forma de avaliação, à medida em que as perguntas são respondidas, a ferramenta apresenta tanto uma pontuação por letra/princípio, como, de forma mais visual, o preenchimento da própria letra. Ou seja, quanto mais "azul" cada letra do acrônimo ficar, mais aderentes ao FAIR serão os conjuntos de dados naquela respectiva dimensão. Ao final, a ferramenta também apresenta uma pontuação geral – o FAIRness, sendo calculada a partir da média das pontuações associadas a cada princípio.

No início da análise, percebeu-se que os três conjuntos de dados depositados no COVID-19 Data Sharing/BR são disponibilizados da mesma forma, seguindo a mesma estrutura. Por esse motivo, esses conjuntos, ao final da análise, receberam pontuações análogas. Desta forma, na apresentação e análise dos resultados, optou-se por apresentar e analisar os resultados obtidos de apenas um conjunto, considerando que são semelhantes. A análise foi realizada em conjunto e de forma simultânea pelos autores objetivando a avaliação sob diversas perspectivas.

Ressalta-se que, para a realização da análise, fez-se necessário o *download* dos conjuntos de dados, como também, o acesso aos metadados dos conjuntos, disponíveis no respectivo repositório, na opção de "Registro Completo". Também acentua-se, que não há apontamentos de alterações ou atualizações dos conjuntos de dados em seus respectivos registros completos, dessa forma, fez-se necessário a identificação das datas referentes a essas alterações ou atualizações, por meio dos metadados inerentes aos arquivos transferidos.

4. Apresentação e análise dos resultados

A primeira seção, correspondente ao princípio *Findable*, é composta por três perguntas. A primeira pergunta questiona sobre o fornecimento suficiente de me-

5 Disponível em: https://satifyd.dans.knaw.nl/. Acesso em 25 de fevereiro de 2021.

6 Disponível em: https://easy.dans.knaw.nl/ui/home. Acesso em 25 de fevereiro de 2021.

tadados. À ferramenta apresenta um ícone **"i"** próximo as perguntas com a apresentação de um texto informativo para cada uma delas, nessa pergunta é apresentada uma lista com 13 itens que indica os parâmetros a serem atendidos quando se busca metadados suficientes. Em meio à análise, percebeu-se que os metadados dos conjuntos de dados analisados não contemplam quatro itens dessa lista, a saber: pessoas que contribuíram para os conjuntos de dados, grupo-alvo dos conjuntos de dados, licença indicando até que ponto os dados são acessíveis e cobertura espacial (localização geográfica em que a pesquisa foi realizada). Ressalta-se que não há informações das pessoas que contribuíram nas pesquisas que originaram os conjuntos de dados, apenas mencionam, como autores, as instituições colaboradoras. Quanto à licença de uso, o repositório menciona em sua página inicial que todos os conjuntos de dados adotam a licença Creative Commons CC-BY de dados abertos, mas considerando que alguns usuários podem ir diretamente para o endereço dos conjuntos de dados, por meio dos seus identificadores persistentes, o mesmo poderá não ter acesso à informação sobre a licença adotada.

A segunda pergunta questiona sobre o uso de padrões como vocabulários controlados, taxonomias (tesauros) ou ontologias para descrição dos conjuntos que, conforme analisado, os conjuntos de dados em questão não proviam pistas acerca do uso de vocabulários controlados. Percebe-se que o não uso de recursos/instrumentos de controle terminológicos ocasionam um déficit, tanto no momento do encontro dos conjuntos dados, quanto na garantia da sua reutilização por outros pesquisadores.

A terceira pergunta indaga sobre o fornecimento de documentação adicional como, por exemplo, um arquivo do tipo README. Apesar de não ter sido encontrado com a mesma nomenclatura, ressalta-se a evidente preocupação do repositório em elaborar um dicionário de dados para cada um dos conjuntos de dados ali publicados. Entende-se que esses dicionários de dados se configuram como uma documentação adicional que descrevem e explicam a forma como os dados estão estruturados, possibilitando que, qualquer pesquisador e/ou instituição, que tiver acesso aos mesmos, os compreenda e os reutilize em suas posteriores investigações.

Dessa forma, quanto ao princípio *Findable*, os conjuntos de dados alcançaram a pontuação FAIRness de 38%.

A segunda seção, correspondente ao princípio *Accessible*, também é composta por três perguntas (quatro, cinco e seis). A quarta pergunta questiona se os metadados são acessíveis ao público mesmo se os dados não estiverem mais disponíveis. Como não foram encontradas informações acerca desta possibilidade de acesso aos metadados, mesmo quando os dados não estiverem mais disponíveis, optou-se por selecionar a opção "não consigo encontrar esta informação".

A quinta pergunta questiona se os conjuntos de dados contêm dados pessoais. De acordo com a Lei Nº 13.709/2018, Lei Geral de Proteção de Dados – LGDP, dados pessoais são informações relacionadas a pessoas naturais identificadas ou identificáveis. Como os dados que compõem os conjuntos de dados analisados foram anonimizados, e enquadram-se na classificação da LGDP de dados anonimizados que são relativos a um titular que não pode ser identificado (LGPD, 2018), respondeu-se a esta questão de forma negativa.

A sexta pergunta questiona qual das licenças de uso foi escolhida de forma a cumprir os direitos de acesso. Como mencionado anteriormente, o repositório informa, em sua página inicial, a licença de uso atribuída a todos os seus conjuntos de dados, mas que, em contrapartida, a licença não é informada nos próprios metadados dos conjuntos. Ressalta-se que, dentre as opções de respostas disponíveis, têm-se: acesso aberto a todos, acesso aberto para usuários cadastrados, acesso restrito por meio de solicitação de autorização, acesso restrito a grupos específicos e outros acessos. Decidiu-se pela opção de acesso aberto para usuários cadastrados, visto que, para *download* dos conjuntos de dados, os usuários precisam fazer um breve cadastro (nome, *e-mail* e instituição), além de concordarem com o termo de responsabilidade que versa sobre o uso ético dos dados e a responsabilidade em atribuir o devido crédito a esses conjuntos por meio de citações.

Observou-se que os registros completos dos conjuntos de dados não expõem metadados que possibilitem o contato entre os usuários para com os detentores dos dados. Evidencia-se que, esse contato pode ser necessário para o esclarecimento de futuras e eventuais questões.

Dessa forma, quanto ao princípio *Accessible*, os conjuntos de dados alcançaram a pontuação FAIRness de 55%.

A terceira seção, correspondente ao princípio *Interoperable*, também é composta por três perguntas (sete, oito e nove). A sétima pergunta questiona se os conjuntos de dados são armazenados em formatos preferenciais. A ferramenta disponibiliza um texto informativo de quais seriam esses formatos preferenciais, em que, para planilhas a ferramenta informa que os formatos preferenciais são ODS e CSV. Os conjuntos de dados analisados podem ser baixados no formato CSV, portanto, todos os dados estão no grupo dos formatos indicados como preferenciais.

A oitava e a nona pergunta, questionam sobre o vínculo a outros (meta)dados e se os mesmos podem ser acessados *online,* e se há o fornecimento de informações contextuais (referência a outros conjuntos, ou a publicações) sobre os conjuntos de dados. Percebeu-se que os conjuntos de dados em questão não possuem informações contextuais ou vínculos a outros (meta)dados agindo de forma contrária ao que o princípio *Interoperable* recomenda.

Dessa forma, quanto ao princípio *Interoperable*, os conjuntos de dados alcançaram a pontuação FAIRness de 50%.

A quarta seção, correspondente ao princípio *Reusable* é composta por três perguntas (dez, onze e doze). A décima pergunta aborda se há informações sobre a procedência dos dados, como: origem dos dados, citações para dados reutilizados, descrição do fluxo de trabalho, histórico de processamento e de versão dos dados. Nesta pergunta, selecionou-se apenas a opção de origem dos dados, visto que é informada na descrição dos mesmos. Quanto a citações para dados reutilizados, descrição do fluxo de trabalho e histórico de processamento e versão dos dados, não foram encontradas informações.

É válido destacar a relevância dessas outras informações para conjuntos de dados. A descrição do fluxo de trabalho possibilita que pesquisadores terceiros possuam um maior entendimento de como os dados foram criados, ou reutilizados, por meio das citações à esses conjuntos de dados.

A décima primeira pergunta, aborda novamente a licença de acesso e uso adotada pelos conjuntos de dados, igualmente abordada anteriormente pela sexta questão, com as mesmas opções de resposta, na qual selecionou-se acesso aberto.

A décima segunda pergunta, questiona se os (meta)dados atendem aos padrões de domínio quanto a organização de forma padronizada dos dados. Selecionou-se a opção do uso de padrões de domínio, considerando a forma estruturada e padronizada em que os dados foram organizados.

Dessa forma, quanto ao princípio *Reusable*, os conjuntos de dados alcançaram a pontuação FAIRness de 74%.

Para uma melhor compreensão e visualização das questões e de como foram respondidas, elaborou-se o Quadro 1, com essa respectiva descrição.

Quadro 1 – Perguntas da SATIFYD e respectivas opções de respostas selecionadas.

Princípio	Perguntas	Opção de resposta selecionada
FINDABLE Seção 1	Você forneceu metadados (informações) suficientes sobre seus dados para que outras pessoas os encontrassem, entendessem e reutilizassem?	Campos de metadados obrigatórios e alguns campos adicionais.
	Você usou padrões como vocabulários controlados, taxonomias (tesauros) ou ontologias para descrever seu conjunto de dados?	Não foram utilizados padrões.
	Você forneceu documentação adicional rica e detalhada?	Arquivo README.
ACCESSIBLE Seção 2	Os metadados são acessíveis ao público mesmo se os dados não estiverem mais disponíveis?	Sim.
	Seu conjunto de dados contém dados pessoais?	Não.
	Qual das licenças de uso você escolheu para cumprir os direitos de acesso anexados aos dados?	Acesso aberto (usuários cadastrados).

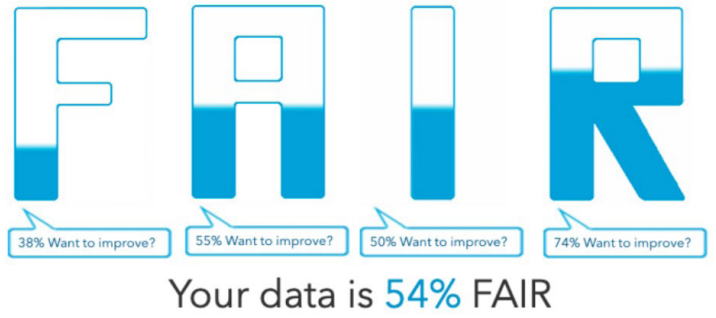
Quadro 1 – Perguntas da SATIFYD e respectivas opções de respostas selecionadas.

Princípio	Perguntas	Opção de resposta selecionada
INTEROPERABLE Seção 3	Os dados em seu conjunto de dados são armazenados em formatos preferenciais?	Todos os dados estão em formados preferenciais.
	Você vincula a outros (meta)dados e esses (meta)dados podem ser acessados *on-line*?	Não.
	Você forneceu informações contextuais sobre seu conjunto de dados?	Sem metadados contextuais.
REUSABLE Seção 4	Que tipo de informação você forneceu sobre a procedência dos seus dados?	Origem dos dados.
	Qual licença de uso você escolheu para seu conjunto de dados?	Acesso aberto (usuários cadastrados).
	Seus (meta)dados atendem aos padrões de domínio?	Padrões de domínio em metadados.

Fonte: traduzido de SATIFYD; dados da pesquisa, 2020.

Ao final, a ferramenta disponibilizou a pontuação final FAIRness de 54%, que se configura como a média das pontuações de cada princípio. Conforme a Figura 1, que ilustra como a ferramenta apresenta o resultado da análise.

Figura 1 – Resultado da análise na SATIFYD.



Fonte: Dados da pesquisa, 2020.

5. Considerações finais

Percebeu-se que, apesar de os conjuntos de dados adotarem algumas práticas propostas pelos Princípios FAIR, o COVID-19 Data Sharing/BR não menciona ou recomenda a adoção dos princípios de forma precedente ao ato da publicação dos dados por parte dos seus detentores (SANTOS; SANT'ANA, 2019).

Como observado, o repositório informa, apenas em sua página inicial, que todos os conjuntos de dados ali publicados adotam a Licença Creative Commons CC-BY de dados abertos e que toda e qualquer publicação ou apresentação que fizer

uso dos dados presentes no repositório devem citar o mesmo. Sugere-se que, todos os conjuntos de dados informem, em seus metadados, a licença de acesso e uso adotada, para que não ocorram equívocos, diante da possibilidade dos usuários acessarem diretamente os conjuntos de dados por meio dos seus identificadores persistentes.

Outro ponto observado foi que, repetidamente, apenas em sua página inicial, o repositório informa que os conjuntos de dados são atualizados periodicamente, devendo ser verificado frequentemente para o *download* de novos dados. Em contrapartida, nos metadados dos conjuntos de dados publicados, não há informações sobre o histórico de versões. Ficando a encargo do usuário verificar, após o *download,* sua data de criação e confirmar se houve uma atualização. Sugere-se que esses históricos de versões sejam informados nos metadados.

Entende-se que quanto mais ferramentas de avaliação estiverem disponíveis para utilização, mais têm-se oportunidades de aprimorar e repensar as formas de avaliação dos conjuntos de dados à luz dos princípios FAIR; a necessidade de maturação da forma avaliativa é inerente ao aprimoramento dela. Fato pertinente também aos repositórios, quanto mais avaliações forem realizadas mais propostas de melhorias serão sugeridas.

Em meio à pesquisa, foram encontradas apenas ferramentas de autoavaliação dos conjuntos de dados. Ou seja, o próprio pesquisador realiza a autoavaliação dos seus conjuntos antes da publicação dos mesmos em repositórios. Desta forma, ressalta-se a importância da criação de ferramentas que possibilitem a análise de conjuntos de dados já publicados. De forma a investigar como está sendo a adoção dos Princípios FAIR por toda a comunidade acadêmico-científica.

Uma iniciativa como a do COVID-19 Data Sharing/BR, que nesse caso, foi desenvolvida diante de uma realidade pandêmica é muito bem vinda, a fundamentação da ideia e a importância da colaboração de instituições como a FAPESP, a Universidade de São Paulo (USP), e a participação do Instituto Fleury, além dos hospitais Sírio-Libanês e Israelita Albert Einstein, demonstram a importância da disponibilização de dados em questões relacionadas com a saúde pública. Recomenda-se, contudo, que mais esforços sejam dedicados para ampliar o nível de aderência aos Princípios FAIR dos conjuntos de dados depositados neste repositório.

6. Referências

ALMEIDA, B. de A. *et al*. Preservação da privacidade no enfrentamento da COVID-19: dados pessoais e a pandemia global. **Ciência & Saúde Coletiva**, v. 25, p. 2487-2492, 2020. Disponível em: https://doi.org/10.1590/1413-81232020256.1.11792020. Acesso em: 15 ago. 2020.

BRASIL. Lei nº 13.709 de 14 de agosto de 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 10 out. 2020.

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

FAPESP. FAPESP COVID-19 **Data Sharing/**BR. *Online*. Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/. Acesso em: 10 ago. 2020.

GO FAIR. FAIR **Principles**. *Online*. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 10 ago. 2020.

SALES, L. et al. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. Disponível em: https://doi.org/10.1162/dint_a_00046. Acesso em: 20 ago. 2020.

LEONELLI, S. Data-from objects to assets. **Nature**, v. 574, p. 317 - 320, 2019. Disponível em: http://dx.doi.org/10.1038/d41586-019-03062-w. Acesso em: 20 ago. 2020.

RICHARDSON, R. J. **Pesquisa Social:** Métodos e Técnicas. 4 ed. São Paulo: Atlas, 2017. 424p.

RODRIGUEZ-IGLESIAS, A. *et al*. Publishing FAIR Data: An Exemplar Methodology Utilizing PHI-Base. **Frontiers in plant Science**. v.7. 2016. Disponível em: https://doi.org/10.3389/fpls.2016.00641. Acesso em: 15 ago. 2020.

SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Camadas de Representação de Dados e suas Especificidades no Cenário Científico. In: DIAS, A. D.; OLIVEIRA, B. M. J. F (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: Editora UFPB, 2019, p. 53-66.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 15 ago. 2020.

WISE, J. *et al*. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. **Drug Discovery Today**, v. 24, n. 4, 2019, p. 933-938. Disponível em: https://doi.org/10.1016/j.drudis.2019.01.008. Acesso em: 15 ago. 2020.

Como citar com o DOI individual

SIMÕES, Anderson Rafael Castro; ANJOS, Renata Lemos dos; DIAS, Guilherme Ataíde Análise dos conjuntos de dados disponíveis no repositório COVID-19 Data Sharing/BR à luz dos princípios FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 91 - 102. DOI: 10.22477/9786589167242. cap7

Princípios FAIR e Linked Data: publicação de cadernos abertos de pesquisa

Luciana Candida da Silva[1] e José Eduardo Santarem Segundo[2]

1. Introdução

Vivencia-se um momento de transformações no modo de fazer e comunicar a pesquisa científica. Estas transformações são ocasionadas pelos avanços na tecnologia, avalanche de produção de dados, pressões de financiamento e cultura colaborativa entre pesquisadores.

Essa nova forma de fazer ciência contempla a divulgação de dados primários, preferencialmente, na medida em que são gerados, e não apenas os casos de sucesso ou de resultados consolidados. A tendência é permitir a colaboração simultânea de modo aberto à ampla contribuição com o propósito de alcançar novos resultados.

Os cadernos de pesquisa são instrumentos de anotações de dados, geralmente experimentais, produzidos em laboratórios para fundamentar as publicações científicas, que são divulgadas normalmente em suas configurações finais. Para Schnell (2015) o caderno de laboratório registra as hipóteses, experimentos e análises iniciais ou interpretações dos experimentos; e serve como um registro legal de propriedades das ideias e resultados obtidos por um cientista. Nota-se que apesar da importância atribuída aos dados registrados em cadernos de pesquisa, estes, em sua maioria, não são divulgados ou não são publicados de forma estruturada de modo a serem reutilizados por usuários humanos e máquinas.

Neste contexto, destacam-se as tecnologias da Web Semântica e Princípios FAIR para orientar na estruturação dos cadernos abertos de pesquisa de modo a torná-los encontráveis, acessíveis, interoperáveis e reutilizáveis, além de aprimorar a capacidade das máquinas de encontrar e usar automaticamente os dados e apoiar a sua reutilização por indivíduos. Sendo assim, este estudo apresenta elementos se-

1 Doutora em Ciência da Informação, Universidade Federal de Goiás (UFG), luciana_candida@ufg.br

2 Livre-Docente em Ciência da Informação, Universidade de São Paulo (USP), santarem@usp.br

mânticos para publicação de cadernos abertos de pesquisa a partir da aplicação dos princípios FAIR, na perspectiva da Web Semântica e *Linked Data* a fim de apoiar as novas práticas científicas.

Este estudo é parte do resultado de uma pesquisa de doutorado que teve como objetivo propor diretrizes semânticas para estruturação e publicação de dados abertos de cadernos de pesquisa, visando melhorias na recuperação e compartilhamento de dados.

2. Cadernos abertos de pesquisa

O termo Caderno Aberto de Pesquisa ou *Open Notebook Science*, foi cunhado por Jean-Claude Bradley, em setembro de 2006, para promover debates sobre a colaboração aberta na ciência e desenvolver técnicas de pesquisas mais eficazes. Para Bradley (2010) o objetivo dos cadernos abertos de pesquisa é tornar os detalhes dos experimentos feitos em laboratórios disponíveis gratuitamente na Web, o que não restringe apenas aos dados bem sucedidos.

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

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

Segundo Schapira e Harding (2019) o caderno aberto de pesquisa inclui vários procedimentos realizados durante uma pesquisa experimental de modo a garantir a replicação bem sucedida dos resultados. Esses conjuntos de procedimentos são mencionados na literatura como protocolo de pesquisa, o qual se configura como principal objeto de publicação e compartilhamento de dados, pois possui uma descrição completa dos procedimentos realizados, dos equipamentos adotados e dos reagentes utilizados durante a pesquisa, além de declarar os objetivos pretendidos,

a discussão dos dados encontrados na pesquisa e registra os resultados alcançados sejam parciais ou finais. Os protocolos devem ser acompanhados de documentos textuais e planilhas, caso seja necessária para a interpretação dos procedimentos de materialização dos objetos de estudo como traços, pontos, gráficos, mapas, espectros e outros.

A proposta de abertura de conjuntos de dados de pesquisa registrados em cadernos de laboratório faz parte de um movimento maior da Ciência Aberta denominado *e-Science,* caracterizado pelo uso intensivo de tecnologias e esforços colaborativos, os quais trazem a oportunidade de se pensar os novos contextos e práticas científicas. Neste contexto, a Foster (2018), classifica os cadernos de pesquisa como parte integrante da terceira dimensão *Open Reproducible Research*, da taxonomia da Ciência aberta, a qual se constitui no ato de oferecer aos usuários livre acesso a elementos experimentais para permitir a reprodução da pesquisa, independente de seus resultados.

3. Princípios FAIR

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

Esta discussão resultou-se em quatro relevantes princípios, com práticas orientadoras para publicação de dados que fossem facilmente encontráveis, acessíveis, interoperáveis e reutilizáveis, por máquinas e humanos, frente a grande quantidade de informações geradas pela ciência contemporânea intensiva em dados. Estes princípios incorporam características que definem que os recursos, ferramentas, vocabulários e infraestrutura de dados contemporâneos devessem ser exibidos para auxiliar na descoberta e reutilização de terceiros (FORCE11, 2014).

Segundo Wilkinson *et al*. (2016), os elementos dos princípios FAIR estão relacionados, mas são independentes e separáveis e podem ser implementados em qualquer combinação, de forma incremental, à medida que os provedores de dados evoluem suas estruturas no sentido de atingir um grau maior dentro do propósito dos princípios FAIR. Os autores esclarecem que estes princípios precedem as escolhas de implementação e não engessam tecnologias para implementação. Sendo assim, este estudo associa os princípios FAIR as tecnologias da Web Semântica e *Linked Data*.

4. Web semântica e Linked Data

A Web Semântica teve inicio em 2001, por Tim Berners-Lee com a colaboração de Hendler e Lassila, a partir da proposta de definir uma maneira eficiente para representar dados na Web e proporcionar melhorias na qualidade da recuperação da informação, permitindo, de acordo com Santarém Segundo (2012, p.106) "aos usuários obter resultados mais precisos e com informação mais próxima do que realmente necessitam".

O termo *Linked Data* se apresenta como princípios para implementação das tecnologias da Web Semântica para publicar e promover a ligação de dados de diferentes fontes na Web, de forma a proporcionar benefícios aos dados. Estes princípios são: 1– usar *Uniform Resource Identifier* (URIs) com nomes para coisas; 2–usar URI HTTP, para que as pessoas possam procurar esses nomes; 3– quando alguém procurar um URI, fornecer informações úteis, usando padrões, como *Resource Description Framework* (RDF) e SPARQL; e 4– incluir *links* para outros URIs, para que os itens relacionados possam ser descobertos (BERNERS-LEE, 2006).

O consórcio World Wide Web (W3C) recomenda um conjunto de tecnologias para publicação de dados abertos e conectados na Web, segundo os princípios do *Linked Data*. Dentre as tecnologias destacam-se o *Resource Description Framework* (RDF) e suas serializações, sendo o RDF um modelo padrão adotado para a descrição de informações estruturadas na Web, permitindo representar a informação de um recurso de forma legível por máquinas (W3C, 2014).

O W3C recomenda um conjunto de 35 (trinta e cinco) Melhores Práticas (MP) para Dados na Web, (DWBP, do inglês *Data on the Web Best Practices*), para melhorar a coerência entre provedor e consumidor, incentivar e permitir a expansão continuada da Web como um meio para o intercâmbio de dados e promover a reutilização de dados de forma confiável.

As trinta e cinco MP para publicar dados na Web são distribuídas em categorias e para cada melhor prática obtém-se um conjunto de benefícios, conforme apresentados no quadro 01.

Quadro 01 - Melhores Práticas e Benefícios

Categoria	Melhor Prática	Benefícios
Metadados	MP 1 - Fornecer metadados para usuários humanos e máquinas	Reuso, compreensão, descoberta e processabilidade
	MP 2 - Fornecer metadados descritivos	Reuso, compreensão e descoberta
	MP 3 - Fornecer metadados estruturados	Reuso, compreensão e processabilidade
Licenças	MP 4 - Fornecer informações de licença de dados	Reuso e confiabilidade
Proveniência de dados	MP 5 - Fornecer informações de proveniência de dados	Reuso, compreensão e confiabilidade
Qualidade de dados	MP 6 - Disponibilizar informações sobre a qualidade de dados e adequações necessárias	Reuso e confiabilidade

Quadro 01 - Melhores Práticas e Benefícios

Categoria	Melhor Prática	Benefícios
Versão de dados	MP 7 - Atribuir versão para cada conjunto de dados	Reuso e confiabilidade
	MP 8 - Fornecer um histórico de versão	Reuso e confiabilidade
Identificadores de dados	MP 9 - Usar URIs persistentes como identificadores de conjunto de dados	Reuso, interligação, descoberta e interoperabilidade
	MP 10 - Usar URIs persistentes como identificadores dentro de conjuntos de dados	Reuso, interligação, descoberta e interoperabilidade
	MP 11 - Atribuir URIs a versões de conjuntos de dados	Reuso, descoberta e confiança
Formatos de dados	MP 12 - Usar formatos de dados padronizados	Reuso e processabilidade
	MP 13 - Usar representações de dados neutras à localidades	Reuso e compreensão
	MP 14 - Fornecer dados em vários formatos	Reuso e processabilidade
Vocabulários de dados	MP 15 - Reutilizar vocabulários, de preferência padronizados	Reuso, processabilidade, compreensão, confiança e interoperabilidade
	MP 16 - Escolher o nível de formalização correto	Reuso, compreensão e interoperabilidade
Acesso a dados	MP 17 - Permitir o acesso completo (em massa)	Reuso e acesso
	MP 18 - Permitir o acesso parcial ao conjunto de dados	Reuso, acesso, interligação e processabilidade
	MP 19 - Disponibilizar dados em vários formatos	Reuso e acesso
	MP 20 - Permitir o acesso em tempo real	Reuso e acesso
	MP 21 - Fornecer dados atualizados	Reuso e acesso
	MP 22 - Explicar os motivos de quando os dados não estiverem mais disponíveis	Reuso e confiabilidade
	MP 23 - Disponibilizar dados através de uma API	Reuso, processabilidade, interoperabilidade e acesso
	MP 24 - Usar padrões da Web como base de APIs	Reuso, interligação, interoperabilidade, descoberta, acesso e processabilidade
	MP 25 - Fornecer documentação à medida que adicionar ou alterar uma API	Reuso e confiabilidade
	MP 26 - Evitar quebrar alterações na sua API	Reuso e interoperabilidade
Preservação de dados	MP 27 - Preservar o identificador e fornecer informações sobre o recurso arquivado	Reuso e confiabilidade
	MP 28 - Avaliar a cobertura de um conjunto de dados antes da sua preservação	Reuso e confiabilidade
Feedback	MP 29 - Coletar *feedback* dos consumidores	Reuso, compreensão e confiabilidade
	MP 30 - Disponibilizar publicamente o *feedback*	Reuso e confiabilidade
Enriquecimento de dados	MP 31 - Enriquecer dados gerando novos dados	Reuso, compreensão, confiabilidade e processabilidade
	MP 32 - Oferecer apresentações complementares	Reuso, compreensão, acesso e confiabilidade
Republicação	MP 33 - Fornecer *feedback* ao publicador original	Reuso, interoperabilidade e confiabilidade
	MP 34 - Seguir os termos de licença	Reuso e confiabilidade
	MP 35 - Citar a publicação original	Reuso, descoberta e confiabilidade

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

Após a apresentação destas MP é possível analisar os elementos semânticos mapeados para descrever os cadernos abertos de pesquisa quanto ao alcance dos dados serem encontráveis, acessíveis, interoperáveis e reutilizáveis, a partir da apli-

cação dos princípios FAIR, das tecnologias da Web Semântica e dos conceitos do *Linked Data*.

5. Elementos semânticos de descrição dos cadernos abertos de pesquisa

Apresenta o resultado do mapeamento de metadados e de vocabulários que visam descrever e individualizar os objetos que compõem o ecossistema dos cadernos de pesquisa, especialmente no que se refere a pesquisas experimentais para estruturação e publicação de cadernos abertos de pesquisa. O mapeamento apresenta um conjunto de elementos semânticos, conforme apresentado no quadro 2.

Quadro 02 - Mapeamento de metadados e propriedades de vocabulários

Planilha de Metadados (Rótulos)	Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets*
Identificador do registro	schema:identifier
Data e horário do registro	schema:dateTime
Autor ▲	schema:author
• Data de nascimento	schema:birthDate
• Data de morte	schema:deathDate
• Profissão/Ocupação	schema:hasOccupation
• Instituição vinculada ▲	schema:memberOf
• Departamento da Instituição	schema:department
Contribuinte /Colaborador ▲	schema:participant
• Data de nascimento	schema:birthDate
• Data de morte	schema:deathDate
• Profissão/Ocupação	schema:hasOccupation
• Instituição vinculada ▲	schema:memberOf
• Departamento da Instituição	schema:department
Instituição ▲	schema:sourceOrganization
Agência de Fomento ▲	schema:funder
• Identificador do agente	schema:Identifier
• Ponto de acesso controlado	skos:prefLabel
• Ponto de acesso variante	skos:altLabel
• Campo de atividade	rdaa:P50387
• Idioma ▲	schema:inLanguage
• Informação de contato	schema:email
Título	schema:name
Subtítulo	schema:alternativeHeadline
Idioma ▲	schema:inLanguage
Formato ▲	dct:format
Tipo ▲	dct:type
Número total de páginas	schema:pagination

Quadro 02 - Mapeamento de metadados e propriedades de vocabulários

Planilha de Metadados (Rótulos)	Propriedade dos Vocabulários Schema.org, DC Terms, SKOS e RDA *Element Sets*
Cobertura espacial ▲	schema:location
Período de execução da pesquisa	schema:startTime
Público	schema:audience
Objetivo pretendido	schema:description
Resultados alcançados	schema:result
Assunto ▲	schema:about
• Identificador ▲	schema:propertyID
• Ponto de acesso controlado	skos:prefLabel
• Ponto de acesso variante	skos:altLabel
• Assunto mais amplo	skos:broader
• Assunto mais específico	skos:narrower
Descrição	schema:description
• Reagentes ▲	schema:activeIngredient
• InChIKey ▲	schema:identifier
• Equipamentos	schema:instrument
• Fórmula molecular ▲	schema:identifier
• Peso molecular ▲	schema:weight
• Técnica de medição	schema:measurementTechique
• Nomes químicos ▲	skos:related
• Nome comercial ▲	skos:related
• Data de criação	schema:dateCreated
• Data de modificação	schema:dateModifield
• Período de encerramento da pesquisa	schema:endTime
• Status da ação	schema:actionStatus
• Error	schema:error
Fonte de dados	schema:provider
Declaração de proveniência	dct:provenance
Licença de uso	schema:license
Declaração de direitos	dct:RightsStatement
Titular dos direitos	dct:rightsHolder
Tamanho do aplicativo	schema:fileSize
Software necessário	schema:availableOnDevice
Registros de exibição	schema:RegisterAction
Controle de uso e usuários	schema:userInteractionalCount
Tipo de interação do usuário	schema:interactionType

Fonte: Silva (2020).

Na primeira coluna encontram-se os metadados identificados por meio de modelagem de dados, e inserção de atributos que descrevem as especificidades dos cadernos de pesquisa, acrescidos de elementos que podem ser enriquecidos com

vocabulários externos, como data de nascimento e data de morte de determinado autor, quando for o caso. Os elementos sinalizados com um triângulo (▲) indicam a importância do reuso de informações de *datasets* externos, por meio de URI, sempre que possível para evitar ambiguidades, garantir padronização e carregar informações adicionais àquelas requeridas pelos metadados. Na segunda coluna encontram-se as propriedades dos vocabulários Schema.org, DC Terms, SKOS e RDA *Elemet Sets* correspondentes aos metadados da primeira coluna.

A partir desta estrutura apresenta uma análise dos elementos serem considerados FAIR, tendo em vista que os princípios FAIR vêm sendo requeridos pela comunidade acadêmica, em especial pelas agências de fomento, como critério de avaliação para financiamento de pesquisas.

5.1 Análise dos elementos semânticos dos Cadernos de Pesquisa quanto serem FAIR

Discute-se o alcance dos elementos indicados na composição das diretrizes para publicação de cadernos de pesquisa quanto aos dados serem encontráveis, acessíveis, interoperáveis e reutilizáveis, a partir da aplicação dos princípios FAIR, tecnologias da Web Semântica e conceitos do *Linked Data*, recomendadas pelo W3C.

5.1.1 Encontrável *(Findable)*

O princípio *findable* recomenda quatro práticas para que os dados sejam encontráveis, sendo que em F1, primeiro princípio, indica a atribuição de identificadores globalmente exclusivo e persistente e em F3, terceiro princípio, sinaliza que incluam explicitamente os identificadores dos dados que descrevem. Neste estudo, os cadernos de pesquisa foram estruturados a partir do uso de identificadores persistentes para descrever nomes de objetos, pessoas, instituições, lugares e assuntos, por meio da definição de *tags* de metadados.

As recomendações do W3C destacam que a descoberta, o uso e a citação de dados na Web dependem fundamentalmente do uso de URIs HTTP que podem ser consultados na internet. Assim, as MP 9, 10 e 11 sinalizam para o uso de URI persistente para conjunto de dados e URI como identificadores de conjuntos de dados. Nesse sentido, é possível observar o mapeamento de *tags* direcionadas para o uso de identificadores ou URIs persistentes, como identificadores de pessoa por meio de vocabulários como o ORCID e o VIAF (*schema:author*, *schema:funder e schema:sourceOrganization*), indicadores para cobertura espacial utilizando o vocabulário GeoNames (*schema:location*) e indicadores de assunto através dos vocabulários LCSH, MeSH e PubChem que podem ser apresentados, a partir da propriedade *schema:about* e seus desdobramentos, conforme apresentado no quadro 2.

Nesta perspectiva, o segundo princípio *findable* (F2) recomenda que um conjunto de dados deva ser descrito por metadados ricos o suficiente para que, uma vez indexados em um mecanismo de busca possam ser encontrados mesmo sem o seu identificador persistente. As MP 1, 2 e 3 recomendam fornecer metadados descritivos, estruturais e administrativos. Os elementos mapeados descrevem dados relacionados à pesquisa experimental, não pretendeu ser demasiadamente exaustivo e sim descrever as informações consideradas o suficiente para que o pesquisador possa analisar e optar pela reprodução ou repetibilidade dos dados.

Na ocasião de implementação destas diretrizes faz-se necessário oferecer metadados para leitura humana, onde o W3C recomenda fornecer metadados como parte de uma página da Web HTML e como um arquivo de texto separado. Para a interpretação de máquinas, os metadados podem ser fornecidos em um formato de serialização Turtle e JSON, ou podem ser incorporados na página HTML (HTML-RDFA ou JSON-LD), e reutilizar padrões existentes e vocabulários populares. Para os cadernos de pesquisa optou-se pelo uso integrado dos padrões de metadados Schema.org e Dublin Core para possibilitar a descrição detalhada dos valores dos metadados. O Schema.org foi selecionado para descrever os objetos digitais em torno dos cadernos de laboratório por possuir o maior número de propriedades que correspondem com as pesquisas experimentais e o Dublin Core que possui múltiplas propriedades para descrever informações como as proveniências, formatos e tipos dos dados. Além desses, adotou-se o vocabulário SKOS para refinar os valores dos metadados.

A prática recomendada em F4, quarto princípio *findable*, é que os metadados sejam registrados e indexados em mecanismos de busca. A MP 12 colabora com o princípio F4 ao recomendar o uso de formatos padronizados e legíveis por máquinas ao publicar dados na Web. Entre os formatos indicados incluem, mas não limitam, a sintaxe de serialização CSV, XML, HDF5, JDON e RDF como RDF/XML, JSON ou Turtle. A MP 24 recomenda adotar padrões como base das APIs e a MP 35 indica citar o publicador original como forma de facilitar a descoberta. Nestas diretrizes disponibilizou a *tag schema:provider* para indicar a fonte dos dados e proporcionar confiabilidade aos dados.

5.1.2 Acessível *(Accessible)*

Segundo Wilkinson *et al.* (2016) a acessibilidade dos dados está relacionada ao uso de protocolos de comunicação padronizados, abertos e gratuitos, que ofereçam autenticação e acesso aos metadados mesmo quando não estiver mais disponível.

A recomendação do W3C é possibilitar o acesso ao conjunto completo de dados de uma determinada pesquisa. Para o fácil acesso a estes conjuntos de dados,

a MP 17 recomenda que a infraestrutura da Web deva ser implementada de modo a permitir o acesso em massa de um conjunto de dados completo com apenas um pedido, evitando inconsistência no acesso individual de dados ao longo de muitas recuperações, bem como permitir o fornecimento de subconjuntos de dados (MP18), caso os consumidores não precisem do conjunto completo.

A Web oferece acesso usando métodos de protocolo de transferência de hipertexto (HTTP) para *download* simples, em massa, de um arquivo. Ainda que os dados estejam distribuídos em vários URIs, estes podem ser organizados em um modelo de contêiner, através do protocolo de transferência de arquivos, para facilitar o acesso em massa aos dados. A distribuição dos dados em vários arquivos permite a recuperação por meio de uma interface de programação de aplicativos (API), método de recuperação mais sofisticado. As MP 18, 20, 23 e 24 mencionam que uma API é a abordagem mais flexível para servir subconjuntos de dados, pois permite a personalização de quais dados são transferidos e fornece dados em tempo real.

O segundo princípio *Accessible* (A2) orienta que os metadados devam ficar acessíveis, mesmo quando os dados não estão disponíveis, enquanto a W3C orienta, por meio da MP 22, fornecer explicações para dados que não estão disponíveis, informando sobre como podem ser acessados e quem pode acessá-los. Nesse aspecto, as recomendações FAIR e MP podem se complementar disponibilizando metadados, mesmo quando os dados não estiverem disponíveis, e ainda assim oferecer mensagens explicativas.

5.1.3 Interoperável *(Interoperable)*

A interoperabilidade refere-se à capacidade de um sistema se comunicar facilmente com outro. Para este benefício ações como atribuir metadados interligados e padrões Web como base das APIs são necessários. Além disso, para que haja uma linguagem formal é necessário o fornecimento de metadados legíveis por humanos e máquinas (MP 1 e 2), uso de indicadores persistentes (MP 9 e 10) e o reuso de vocabulários padronizados (MP 15 e 16).

Para a estruturação de cadernos de pesquisa selecionou-se vocabulários que descrevem seus propósitos, como a descrição de autoridades com refinamento de atributos que indicassem as datas, a profissão, o campo de atividade, os meios de comunicação com os agentes e a possiblidade de vincular os pesquisadores às instituições e departamento pelos quais fazem parte. Além disso, buscaram-se vocabulários que descrevessem as especificidades de uma pesquisa experimental como nome de compostos químicos, propriedades químicas, procedimentos realizados durante a pesquisa, períodos de realização, *status* da pesquisa e uma maneira de informar se a pesquisa foi bem sucedida ou não; buscou também vocabulários que

possibilitassem a descrição de proveniências dos dados. Após o detalhamento dos atributos que descrevem as especificidades dos cadernos de pesquisa optou-se pelos vocabulários Schema.org, DCTerms e SKOS, além dos vocabulários de valores como GeoNames, VIAF, ORCID dentre outros exemplificados em F1 a F3. Como modelo de representação, a prática recomendada é o uso do RDF e suas serializações, também mencionada em F2. Ademais, o uso de vocabulários amplamente reconhecidos favorece os benefícios interoperabilidade, processabilidade, compreensão, confiabilidade e o reuso dos dados.

5.1.4 Reutilizável *(Re-Usable)*

Este princípio é especialmente importante para o contexto dos cadernos de pesquisa, pois reflete a aplicação dos princípios anteriores (encontrável, acessível e interoperável) e ao objetivo da estruturação que é o reuso dos dados por pesquisadores em novas pesquisas.

O princípio R1 estabelece que os metadados devam ser descritos com pluralidade de atributos precisos e relevantes, sendo assim, foram mapeados metadados de proveniência, de descrição das especificidades dos cadernos de pesquisa, metadados administrativos e de uso. Para descrever os valores textuais, o uso de identificadores persistentes é recomendado para fins de enriquecimento de dados.

O princípio R1.1, a MP4 e a MP34 destacam a importância de fornecer um tipo de licença para evitar limitações de reuso e formalizar legalmente a disponibilização de dados para reutilização do trabalho de outra pessoa. O princípio R1.2 e MP 5 recomendam que os metadados devem estar associados à sua proveniência. Nestas diretrizes foram mapeadas as *tags* para fonte dos dados (*schema:provider*) para indicar a origem dos dados, declaração de proveniência (*dct:provenance*) para explicitar alterações na propriedade e custódia de um objeto, licença (*schema:license*) para permitir fazer o uso do objeto, declaração de direitos (*dct:RigtsStatement*), titular dos direitos (*dct:rightsHolder*) para indicar o nome do agente que possui ou gerencia os direitos do objeto, data de criação (*schema:dateCreated*) e modificação dos dados (*schema:dateModifield*) para informar a data original e das modificações.

A MP 13 orienta disponibilizar informações sobre parâmetro de localidade, para evitar dificuldades de compreensão dos dados que se modificam de um idioma para outro, como por exemplo, informar o idioma na *tag schema:inLanguage* do padrão Schema.org.

A MP 14 recomenda sempre que possível à disponibilidade de dados em múltiplos formatos, quando mais de um se adequa ao seu uso pretendido. Ao implementar estas diretrizes recomenda-se a conversão dos dados para os formatos RDF/XML, JSON e Turtle, os quais atendem a recomendação das MP 12 de uso de formato

legível por máquina, MP 14 formatos em múltiplos formatos, MP 15 vocabulários conhecidos e padronizados.

6. Considerações finais

O fornecimento de metadados administrativos, descritivos, de proveniência, de preservação e de uso, com a indicação de implementação a partir de vocabulários padronizados e que permitam a leitura por pessoas e máquinas, além da demasiada recomendação da atribuição de URIs para indicação de nomes e enriquecimento de dados, podem garantir que dados de pesquisa de cadernos de laboratórios sejam encontráveis.

Seguir as recomendações de uso de APIs favorecerá a acessibilidade aos dados. Além destes elementos recomenda-se o uso de explicação contextual sobre a situação dos dados e seus metadados, de forma a proporcionar interoperabilidade dos dados. Destaca-se que unificar as recomendações anteriores à indicação de licença de uso, preferencialmente, de domínio público, bem como o detalhamento de metadados associados a proveniência dos dados, facilita o reuso dos dados de pesquisa dos cadernos de laboratórios publicados a partir destas diretrizes. Vale salientar que nem todas as recomendações relacionadas aos princípios da acessibilidade estão contempladas nestas diretrizes de estruturação, no entanto fica a recomendação para a ocasião de implementação.

7. Referências

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, 17 may 2001.

BERNERS-LEE, T. **Linked Data**. 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 13 jun. 2018.

BRADLEY, J. C. The impact of open notebook Science. **Information Today**, [S.l], v. 27, n.8, p. 50-51, set. 2010. Entrevista realizada por Richard Poyder. Disponível em: http://www.infotoday.com/it/sep10/Poynder.shtml#top. Acesso em: 02 set. 2019.

FORCE11. The Future of Research Communications and e-Scholarship. **Guiding principles for findable, accessible, interoperable and reusable data publishing version B1.0**. 2014. Texto digital. Disponível em: https://www.force11.org/fairprinciples. Acesso em: 13 jun. 2018.

FOSTER. **Open reproducible research**. 2018. Disponível em: https://www.fosteropenscience.eu/taxonomy/term/102. Acesso em: 02 ago. 2020.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web best practices**.

W3C, 2017. Texto digital. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 20 mar. 2020.

RAUTENBERG, S.; SOUZA, L.; DALL'AGNOL, J. M. H.; MICHELON, G. A. **Guia prático para publicação de dados abertos na Web**. Curitiba: Appris, 2018. 280 p.

SANTAREM SEGUNDO, J. E. Tim Berners-Lee e a ciência da informação: do hipertexto à Web Semântica. *In*: SANTAREM SEGUNDO, J. E.; SILVA, M. R.; MOSTAFA, S. P. (org.). **Os pensadores e a Ciência da Informação**. Rio de Janeiro: E-papers, 2012. p. 101-109.

SCHAPIRA, M.; HARDING, R. J. Open laboratory notebooks: good for Science, good for society, good for scientists. **F1000Research Open for Science**, 2019. Disponível em: https://doi.org/10.12688/f1000research.17710.1. Acesso em: 21 set. 2019.

SCHNELL, S. Ten Simple Rules for a computational biologist's laboratory notebook. **PloS Computational Biology**. São Francisco, v.11, n.9, 2015. Disponível em: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004385. Acesso em: 28 fev. 2021.

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

SILVA, L. C.; SANTAREM SEGUNDO, J. E.; SILVA, M. F. Princípios FAIR e melhores práticas do Linked Data na publicação de dados de pesquisa. **Informação&Tecnologia** (ITEC), Marília/João Pessoa, v.5, n.2, p.81-103, jul./dez. 2018.

W3C. World Wide Web Consortium. RDF - **Resource Description Framework**. 2014. Disponível em: https://www.w3.org/RDF/. Acesso em: 27 mar. 2020.

WILKINSON, M. *et al*. The FAIR guiding principles for scientific data management and stewardship. **Sci Data**, n. 3, 2016. DOI: 10.1038/sdata.2016.18

Como citar com o DOI individual

SILVA, Luciana Candida da; SANTAREM SEGUNDO, José Eduardo. Princípios FAIR e Linked Data: publicação de cadernos abertos de pesquisa. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 103-116. DOI: 10.22477/9786589167242.cap8

Implementação dos princípios FAIR em repositórios de dados científicos: uma análise comparativa das infraestruturas de software do DSpace e Dataverse

Fabiano Couto Corrêa da Silva[1] e Marcello Mundim Rodrigues[2]

1. Introdução

Ao planejar políticas institucionais voltadas à gestão e curadoria de dados científicos, Instituições de Ensino Superior se deparam com inúmeras questões de respostas complexas, por exemplo, a identificação da infraestrutura de *software* mais apropriada à preservação e organização de dados científicos heterogêneos. Da mesma maneira, os programas e organizações que financiam, conduzem ou de outra forma apoiam pesquisas são responsáveis por promover e garantir uma adequada gestão de dados e informações provenientes de suas atividades. Por mais óbvio que seja, vale reforçar que um eficaz gerenciamento de dados se torna essencial para garantir que tais ativos estejam acessíveis, agora e no futuro.

Os Princípios FAIR (dados encontráveis, acessíveis, interoperáveis, e reusáveis) foram desenvolvidos para ajudar a resolver barreiras comuns à descoberta e reutilização de dados, há muito reconhecido como um problema na pesquisa científica, apresentando diretrizes que apoiam a retenção e disponibilidade em longo prazo desses conjuntos de dados. Os aspectos FAIR de dados localizáveis e acessíveis estão principalmente relacionados ao local onde os dados são depositados. Pontos importantes a serem levados em consideração incluem a disponibilidade de identificador persistente de objeto digital, metadados, monitoramento de reutilização de dados, licenciamento, controle de acesso, retenção e disponibilidade em longo prazo.

1 Doutor em Información y Documentación pela Universitat de Barcelona, Professor no Programa de Pós-Graduação em Ciência da Informação, Universidade Federal do Rio Grande do Sul. E-mail: fabianocc@gmail.com.

2 Doutorando em Ciência da Informação pela Universidade Federal de Minas Gerais, Bibliotecário-Documentalista na Universidade Federal de Uberlândia. E-mail: marcellomundim@ufu.br.

Os aspectos FAIR de dados interoperáveis e reutilizáveis ressaltam a necessidade de se pensar em questões que abrangem o formato dos dados (proprietário x aberto), sua atualização ou obsolescência, a interoperabilidade (abertura via *Application Programming Interface* [API]) do repositório selecionado a outros meta-repositórios internacionais ou disciplinares, ou outras ferramentas de aprimoramento. O aspecto da documentação detalhada também é levado em consideração na capacidade de reutilização dos dados.

A partir disso, destaca-se que a pesquisa teve como objetivo identificar as diferenças entre duas infraestruturas de *software* voltadas à gestão e curadoria de dados científicos à luz do padrão proposto pelos princípios FAIR.

2. Referencial teórico

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. Embora o acesso aberto lute pelo acesso gratuito sem barreiras, não significa que todas as 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).

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

Lee e Stvilia (2014) 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).

Schopfel e outros (2014) exploram dados científicos relacionados a teses e dissertações eletrônicas como uma parte específica da emergente e-infraestrutura de pesquisa. 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).

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).

Disponibilizar acesso a dados científicos relacionados a teses e dissertações digitais é um desafio a bibliotecas acadêmicas, e com isso, Schopfel e outros (2014) 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 científicos heterogêneos não podem ser comparados ao tipo de *Big Data* produzido pelo CERN e outros, pois são similares a dados pessoais[3]. 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).

Amorim e outros (2016) 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). Para os autores, o acesso ao

3 Por dados pessoais, entende-se que são dados gerados em pequeno volume, porém em diversos formatos.

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, entendem que a existência de uma 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 destacam que 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[4] podem ser instalados e executados 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. 856-857, tradução nossa). O *DSpace*, conhecido por sua capacidade de lidar com publicações de pesquisa, tem sido reconhecido também por manusear dados científicos (AMORIM *et al*., 2016, p. 858, tradução nossa).

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).

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).

Garnett e outros (2017) 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).

4 "O Fedora cria uma plataforma inovadora, livre e de código aberto para *hardware*, nuvens e contêineres que permite desenvolvedores de *software* e membros da comunidade criem soluções personalizadas para seus usuários". (RED HAT, 2020). O Fedora possui sistemas operacionais como soluções específicas a nichos especializados.

Dados FAIR devem ser legíveis por máquina e acionáveis; não são equivalentes a dados abertos; é uma aspiração, nunca é cem por cento FAIR; ao publicar dados restritos, licenças e acordos de uso de dados devem ser claramente definidos pelos autores ou provedores de dados; uma plataforma de repositório como o *Dataverse* pode facilitar muito a criação de dados científicos FAIR; porém, os autores dos dados devem contribuir usando padrões de metadados e vocabulários de comunidade apropriados (CROSAS, 2019, tradução nossa).

Entre as plataformas *DSpace* e *Dataverse*, Rocha e outros (2018) concluem que:

> O *Dataverse* possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais e políticas de gestão distintas para unidades ou grupos, incluindo esquemas de metadados e licenças. Isso é possível em *DSpace*, entretanto, exige adaptações ou configurações, com algumas limitações no controle de versões (ROCHA *et al.*, 2018, p. 74).

Brownlee (2009) afirma que o *Dublin Core* (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).

Garnett e outros (2017) vão além e 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 à 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).

3. Metodologia

O estudo se configura numa pesquisa aplicada, qualitativa, exploratória, descritiva e documental (CRESWELL, 2014). Os objetos de estudo desta investigação foram o *DSpace* e o *Dataverse*, plataformas que se destinam à concepção de ambientes virtuais conhecidos como repositórios digitais. Não se objetivou investigar outras infraestruturas de *software*, visto que as recentes investidas em repositórios institucionais têm se concentrado nessas duas opções. Sendo assim, buscou-se avaliar a conformidade da infraestrutura padrão do *DSpace* e do *Dataverse* com os Princípios FAIR. Desse modo, foi possível identificar se ambas as plataformas

cumprem requisitos mínimos para servirem como instrumento natural à "FAIRificação" (processo de adaptação FAIR) da investigação, ao configurar a descoberta de dados, sua acessibilidade, interoperabilidade e reutilização.

Os princípios FAIR se dividem em subprincípios, cada um correspondendo a um requisito sugerido como de excelência à gestão e curadoria de dados científicos. Identificou-se que os subprincípios referentes à infraestrutura de *software* são F.1, F.3, F.4, A.1, A.1.2, A.2, e R.1.1.

As informações encontradas que serviram aos resultados foram recuperadas por meio da literatura da área e dos documentos oficiais hospedados em seus respectivos *Websites*. A técnica aplicada foi a análise de conteúdo. Para a apresentação dos resultados, criou-se um *check-list* referente aos subprincípios FAIR. Nele, buscou-se responder a seguinte questão: a infraestrutura de *software* (*DSpace* e/ou *Dataverse*) está em conformidade com o subprincípio FAIR x, y, z [...]?

4. Resultados

Os Princípios FAIR sugerem padronização de técnicas e ambientes referentes à gestão e curadoria de conjuntos de dados científicos. A gestão envolve processos de organização do conhecimento; a curadoria, preservação em longo prazo. Preservar significa garantir segurança em função do tempo; organizar, tornar algo encontrável e acessível. Portanto, fazer com que dados e metadados se tornem FAIR exige esforços em diferentes níveis de um mesmo fluxo de trabalho.

Pode-se dizer que a proposição do fluxo de trabalho FAIR se divide em três camadas, a começar: a) camada padrão da infraestrutura do *software* escolhido para armazenar e preservar dados e metadados; b) camada de conhecimento técnico de gestores, curadores e analistas; c) camada de conhecimento de domínio do depositante/dono dos dados e metadados.

A primeira camada possui característica objetiva, pois mesmo com código fonte aberto, as configurações "de fábrica" de um *software* estão à disposição de seus usuários/clientes, garantindo assim mínima padronização de suas funções.

As demais camadas percebidas estão sujeitas ao conhecimento tácito (subjetividade) dos agentes envolvidos em um mesmo projeto de gestão e curadoria de dados científicos, mesmo havendo orientação a partir de uma política institucional bem estabelecida. Por exemplo, quando os subprincípios F.2 e R.1 sugerem que: F.2. Dados são descritos com metadados valiosos; e R.1. (Meta)dados são ricamente descritos com uma pluralidade de atributos relevantes e precisos (GO FAIR, 2019, tradução nossa); eles estão se referindo à qualidade da descrição dos dados pelo uso de metadados e padrões ou esquemas de metadados que são inerentes a funções humanas.

Os dados FAIR podem ser concebidos como um espectro ou continuum variando de objetos digitais parcial a completamente FAIR. Semelhante às cinco estrelas de dados abertos, diferentes níveis FAIR podem ser concebidos para articular condições mínimas à descoberta e reúso de dados FAIR ricamente documentados e funcionalmente vinculados. Isso vai variar de acordo com a comunidade. Alguns dos princípios serão triviais a certos domínios de pesquisa e problemáticos a outros; portanto, cada campo de pesquisa precisa definir o que significa ser FAIR e decidir as medidas apropriadas para avaliar isso (EUROPEAN COMMISSION, 2018, p. 51, tradução nossa).

Portanto, percebe-se que alguns subprincípios FAIR podem ser responsáveis pela institucionalização de políticas de gestão e curadoria de dados científicos discrepantes entre instituições de pesquisa, seja pela divergência entre seus objetivos, necessidades, ou equipes e usuários/clientes. Não se objetivou investigar os subprincípios FAIR que se destinam à orientação de práticas subjetivas. Assim, a pesquisa identificou os subprincípios F.1, F.3, F.4, A.1, A.1.2, A.2 e R1.1 como pertencentes à primeira camada supracitada.

Esses subprincípios dispõem que: F.1. Aos (meta)dados são atribuídos um identificador único e persistente globalmente, 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; A.1. (Meta)dados são recuperáveis por meio de seus identificadores usando um protocolo de comunicação padronizado; 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; e R.1.1. (Meta)dados são liberados com uma clara e acessível licença de uso dos dados (GO FAIR, 2019, tradução nossa).

4.1. *Check-list*

1) A infraestrutura de *software* está em conformidade com o subprincípio FAIR F.1?

DSpace: Sim. O *Handle* é o sistema de identificação persistente padrão no *DSpace* (UNESCO, 2014, tradução nossa).
Dataverse: Sim. A rede *Dataverse* é uma aplicação de código aberto que provê diretrizes e ferramentas para a citação de dados. O *Dataverse* especifica o registro global *Handle* como seu sistema de identificação persistente. O DOI também pode ser utilizado como sistema identificador padrão do *Dataverse* (LEE; STVILIA, 2014, p. 18-19, tradução nossa).

2) A infraestrutura de *software* está em conformidade com o subprincípio FAIR F.3?

DSpace: Sim. O *DSpace* oferece o *Dublin Core* (DC) como esquema de metadados descritivos pré-definido (BROWNLEE, 2009, p. 4, tradução nossa). O DC possui em seus padrões a *tag dc:identifier* à descrição do identificador persistente atribuído a dados e metadados.
Dataverse: Sim. O *Dataverse* permite a citação para conjunto de dados inteiro. DOI, com URL e metadados registrados no *DataCite*. Também, a citação para arquivo de dados, com DOI e URL para cada arquivo (CROSAS, 2019, tradução nossa).

3) A infraestrutura de *software* está em conformidade com o subprincípio FAIR F.4?

DSpace: Sim. Possui mecanismo de pesquisa integrado: o *DSpace* vem com o *Apache Solr*, uma plataforma de pesquisa corporativa de código aberto que permite a pesquisa filtrada (facetada) e a navegação em todos os objetos. O texto completo dos formatos de arquivo comuns é pesquisável, junto com todos os campos de metadados. As interfaces de navegação também são configuráveis (DURASPACE, 2020, tradução nossa). O *DSpace* é indexado no *Google Scholar* (UNESCO, 2014, tradução nossa).
Dataverse: Sim. O *Dataverse* permite citação e metadados detectáveis usando padrões *DataCite*, *schema.org*, *Dublin Core*, DDI, e *Schema.org* JSON-LD (encontrável na *Google Dataset Search*). (CROSAS, 2019, tradução nossa).

4) A infraestrutura de *software* está em conformidade com o subprincípio FAIR A.1?

DSpace: Sim. Um único servidor *Handle* normalmente abre três ouvintes de rede, nas portas 2641 UDP, 2641 TCP e 8000 TCP. A porta 2641 (UDP e TCP) é o número da porta atribuído pela *Internet Assigned Numbers Authority* (IANA) para o protocolo de cabo *Handle*. O modelo de serviço *Handle* e o protocolo de conexão são descritos em RFC 3650, RFC 3651 e RFC 3652. O TCP geralmente é necessário a solicitações administrativas e é usado como reserva para resolução quando o UDP está lento ou indisponível. A porta 8000 oferece uma interface HTTP e HTTPS. Os servidores *Handle* usam "unificação de porta" para que o HTTP e o HTTPS estejam disponíveis na mesma porta. Se as portas do protocolo

Handle padrão não estiverem disponíveis, seus clientes podem recorrer ao tunelamento do protocolo com fio sobre o HTTP. Para qualquer solicitação HTTP que combine o nome de domínio do *proxy* com um *handle*, por ex.: http://hdl.handle.net/20.1000/5555, um dos servidores *proxy* irá consultar o *handle*, obter a URL no registro do *handle* (ou se houver vários URLs no registro do *handle*, ele selecionará um, e essa seleção não está em uma ordem específica) e enviará um redirecionamento HTTP para esse URL ao navegador do usuário. Se não houver nenhum valor de URL, o *proxy* exibirá o registro do *handle* (HANDLE.NET, 2018, tradução nossa).
Dataverse: Sim. Na manutenção do *Handle* como identificador persistente padrão, o processo será igual ao supracitado.

5) A infraestrutura de *software* está em conformidade com o subprincípio FAIR A1.2?

DSpace: Sim. Segurança: o *DSpace* fornece seu próprio sistema integrado de autenticação/autorização, porém também pode se integrar a sistemas de autenticação existentes, como LDAP ou Shibboleth (DURASPACE, 2020, tradução nossa). A distribuição atual do *software Handle.Net* usa as bibliotecas de criptografia padrão *Java* para rotinas de criptografia de baixo nível. O sistema *Handle* fornece duas formas de autenticação: chave pública e chave secreta. Na implementação atual, a autenticação de chave pública é executada usando o algoritmo DSA ou RSA. A autenticação de chave secreta depende de um algoritmo MAC seguro. Em geral, a autenticação de chave secreta usa três partes: (1) o cliente de autenticação; (2) o servidor onde o cliente está executando uma operação; e (3) outro servidor capaz de verificar a autenticação do cliente (HANDLE.NET, 2018, tradução nossa).
Dataverse: Sim. Caso sejam arquivos de dados restritos, autenticação e autorização serão necessárias (CROSAS, 2019, tradução nossa). Na tentativa de acesso a dados restritos via seu *Handle*, o protocolo de comunicação será idêntico ao descrito anteriormente.

6) A infraestrutura de *software* está em conformidade com o subprincípio FAIR A.2?

DSpace: Sim. O *DSpace* é um conjunto de aplicativos *Web* em *Java* e programas utilitários em cooperação que mantêm um armazenamento de ativos e um de metadados associados. Os aplicativos *Web* fornecem interfaces para adminis-

tração, depósito, importação, pesquisa e acesso. O armazenamento de ativos é mantido em um sistema de arquivos ou sistema de armazenamento semelhante. Os metadados (incluindo informações de acesso e configuração) são armazenados em um banco de dados relacional. Ademais, o *DSpace* possibilita o embargo temporário aos dados via solicitação do autor/criador, no entanto, mantém o acesso a seus metadados (DURASPACE, 2020, tradução nossa).
Dataverse: Uma página de destino desativada com os metadados básicos de citação sempre estará acessível ao público se ele usar a URL persistente (*Handle* ou DOI) fornecida na citação para esse conjunto de dados. Os usuários não serão capazes de ver nenhum dos arquivos ou metadados adicionais que estavam disponíveis antes da desativação (DATAVERSE PROJECT, 2020, tradução nossa). O *Dataverse* armazena informações da estrutura do pacote (*dataset*) em base de dados relacional, isto é, armazena pacotes de forma dependente do *software*. Entretanto, o *Dataverse* permite a exportação dos metadados de um *dataset* (arquivos do *dataset* não inclusos) no formato DDI *Codebook*, que resulta em um arquivo XML que descreve todo o pacote, incluindo metadados estruturais (estruturas físicas e lógicas dos documentos, além de variáveis em documentos tabulares). (ROCHA *et al.*, 2018, p. 47).

7) A infraestrutura de *software* está em conformidade com o subprincípio FAIR R1.1?

DSpace: Sim. A *Creative Commons* é a licença padrão no *DSpace* (UNESCO, 2014, tradução nossa).
Dataverse: Sim. Por padrão, todos os novos conjuntos de dados criados por meio da *Web interface* de usuário do *Dataverse* recebem uma Dedicação de Domínio Público da *Creative Commons* CC0 (DATAVERSE PROJECT, 2020, tradução nossa).

5. Considerações finais

A organização e manipulação de dados são o grande desafio para este início de década de 2020. De acordo com a *International Data Corporation* (2020), a cada dois anos dobramos a quantidade de dados produzidos. Na ciência não é muito diferente.

Essa preocupação em disponibilizar é ocasionada ao final da pesquisa, o que faz com que os dados disponibilizados nem sempre atendam aos princípios FAIR, tendo uma baixa, ou nenhuma semântica, com uma diversidade de padrões e formatos. Para que os dados possam ser melhor utilizados, estes deveriam ter uma

rica semântica, e de acordo com Tim Bernes-Lee, estarem classificados como dados cinco estrelas.

Claramente, a infraestrutura é importante, porém para os dados científicos, o acesso e sua reutilização não dependem apenas do desempenho do repositório, mas de características formais associadas com os conjuntos de dados e os processos associados à sua produção. Nesse sentido, ficou demonstrado que ambas as plataformas analisadas (*DSpace* e *Dataverse*) estão em conformidade com os subprincípios FAIR investigados, portanto, são adequadas à gestão e curadoria de dados científicos. No entanto, há que se atentar ao objetivo e política institucional durante implementação de um repositório de dados científicos. Uma organização que possui um repositório institucional (bibliográfico) implementado e que dispende de poucos recursos para investir em um novo projeto (como seria na escolha do *Dataverse*) pode optar por adaptar o *DSpace* ao gerenciamento e curadoria de dados científicos, sem grandes perdas. O *DSpace* permitiria maior controle dos documentos e dos dados conjuntamente, além da facilidade em torná-los ligados (linked data). Por outro lado, a opção do *Dataverse* traria uma plataforma focada unicamente na gestão e curadoria de dados científicos, além de possibilitar uma maior visibilidade dos depósitos institucionais, uma vez que sua infraestrutura permite o compartilhamento de dados científicos entre instituições de ensino superior e pesquisa de todo o globo.

6. REFERÊNCIAS

AMORIM, R. C. *et al*. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Universal Access in the Information Society**, 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.

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.

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.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES. HANDLE.NET

(Ver. 9) Technical Manual. Reston, Virginia, 2018. Disponível em: http://www.handle.net/tech_manual/HN_Tech_Manual_9.pdf. Acesso em: 18 out. 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: 16 out. 2020.

DATAVERSE PROJECT. **Data Management**. Cambridge, MA, 2020. Disponível em: https://guides.dataverse.org/en/latest/user/dataset-management.html. Acesso em: 19 out. 2020.

DOORN, P.; TJALSMA, H. Introduction: archiving research data. **Archival Science**, 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.

DURASPACE. **Technical Specifications**: DSpace. Beaverton, OR, 2020. Disponível em: https://duraspace.org/wp-content/uploads/dspace-files/specsh_dspace.pdf. Acesso em 18 out. 2020.

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 19 out. 2020.

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: 16 out. 2020.

International Data Corporation. Disponível em: https://www.idc.com. Acesso em: 20 out. 2020.

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.

RED HAT. Fedora. [S. l.], 2020. Disponível em: https://getfedora.org/. Acesso em: 03 mar. 2021.

ROCHA, R. P. da *et al*. Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório 2018. Porto Alegre, RS: UFRGS, 2018. Disponível em: http://hdl.handle.net/10183/185126. Acesso em: 19 out. 2020.

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.

Como citar com o DOI individual

SILVA, Fabiano Couto Corrêa da; RODRIGUES, Marcello Mundim. Implementação dos princípios FAIR em repositóriosde dados científicos: uma análise comparativa das infraestruturas de software do DSpace e Dataverse. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 117-128. DOI: 10.22477/9786589167242.cap9

Tecnologias para gestão de dados de pesquisa segundo preceitos FAIR

Milton Shintaku[1], André Luiz Appel[2], Alexandre Faria de Oliveira[3]

1. Introdução

Desde que a ciência surgiu como atividade separada da filosofia, a ciência tem evoluído conforme a comunidade científica vem se estruturando, se adaptando e influenciando a mudança tecnológica. Nesse sentido, a tecnologia sempre esteve presente, sendo um dos resultados decorrentes da ciência, principalmente das ciências rígidas, em parceria com as engenharias. Entretanto, pode-se dizer que a computação e a Web, criadas na metade e no final do século xx, respectivamente, foram marcos que tiveram impacto significativo nas ciências, entre outras atividades humanas. Isto é de tal modo nítido que já aparecem estudos da chamada ciência artificial e ciência virtual, complementando as chamadas ciências naturais, nos quais, muitas vezes, simulam-se fenômenos naturais por meio da informática, ou mesmo criam-se cenários virtuais para pesquisas.

Na comunicação científica, a tecnologia, principalmente a web, também teve impacto nos processos adotados pelos pesquisadores, tanto que, para muitos pensadores, a Web tem o mesmo impacto no compartilhamento da informação que teve a invenção da prensa móvel de Gutenberg. Nesse caminho da Web, ao final do século xx, dois movimentos surgiram e se confirmaram, sendo aceitos na comunidade científica mundial: os Arquivos Abertos (Open Archives) e o Acesso, que, por terem as mesmas siglas em inglês e português, muitas vezes são confundidos, mas que possuem aspectos diferentes.

1 Doutor em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia, shintaku@ibict.br

2 Doutor em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia, andreappel@ibict.br

3 Mestrando em Ciência da Informação pela Universidade Brasília, Instituto Brasileiro de Informação em Ciência e Tecnologia, alexandreoliveira@ibict.br

Os Arquivos Abertos têm alinhamento tecnológico, centrado na interoperabilidade, na possibilidade de troca de informações entre sistemas informatizados, voltados inicialmente para as chamadas bibliotecas digitais. Esse movimento foi criado pela Convenção de Santa Fé, descrito por Van de Sompel e Langoze (2000) como voltado à criação de infraestrutura informacional para a interoperabilidade, principalmente de *preprints*. Logo após, foi utilizado por Suleman e Fox (2001) para a criação de bibliotecas digitais de teses e dissertações, para a criação *da Networked Digital Library of Theses and Dissertations* (NDLTD). No Brasil, os Arquivos Abertos influenciaram o desenvolvimento da Biblioteca Digital de Teses e Dissertações (BDTD).

O Acesso Aberto, por sua vez, tem aspectos mais filosóficos. Nascido da chamada crise dos periódicos, defende o acesso sem restrições a resultados de pesquisa. Para a implementação desse movimento, Harnad *et al.* (2004) sugeriram o uso de revistas de acesso aberto (via dourada) e o uso de repositórios (via verde). No Brasil, iniciativas como *Bioline International*, de 1993, e *Scientific Electronic Library Online* (SciELO), de 1997, foram pioneiras no suporte à publicação de conjuntos de revistas de acesso aberto em meio eletrônico, representando a via dourada. Posteriormente, várias revistas passaram a ofertar acesso gratuito, e universidades criaram repositórios para dar amplo acesso e vazão à sua produção científica.

Mesmo apresentando origens e configurações distintas, os Arquivos Abertos e Acesso Aberto se mesclaram, visto que alguns conjuntos de softwares para criação de revistas e repositórios combinam alguns preceitos dos dois movimentos. Entretanto, durante algum tempo, em universidades brasileiras, bibliotecas de teses e dissertações e repositórios coexistiam, por terem nascido de movimentos diferentes. De certa forma, o Acesso Aberto, por ter aspectos conceituais, englobou os arquivos abertos em muitos casos.

Neste caminho de abertura das atividades da ciência, outro movimento vem se firmando, o chamado Dados Abertos de Pesquisa, que atua na disseminação dos dados coletados ou gerados no âmbito de pesquisas científicas. O compartilhamento de dados por meio da Internet não é novidade, tanto que os dados de governo têm sido dispostos publicamente, como forma de dar maior transparência à administração pública, no caso do Brasil em concordância com a Lei de Acesso à Informação (LAI). Entretanto, dados de natureza sensível, decorrentes de pesquisas, ainda requerem atenção e discussões, da mesma forma que o ato de compartilhar requer implementação de plataformas informatizadas próprias aos diversos tipos e formatos de dados de pesquisa.

Nesse contexto, os Dados Abertos ainda requerem discussões para a sua implementação, mesmo que apresentem vantagens às pesquisas. Um dos pontos en-

contra-se nas tecnologias disponíveis a serem utilizadas para depósito de dados. Outro ponto envolve questões de proteção, restrição, sensibilidade e tantos outras que estão no escopo da gestão de dados de pesquisa, e que precisam de maior consenso para que as tecnologias possam atingir níveis de consenso, explicitados, por exemplo, por iniciativas como os Princípios FAIR.

2. Dados abertos de pesquisa

No modelo de comunicação científica proposto por Björk (2007), questões relacionadas aos dados de pesquisa aparecem nos processos de execução da pesquisa e comunicação dos resultados. Sobre a execução, o autor relata que a pesquisa contempla quatro atividades globais, sendo que uma delas envolve coleta de dados existentes em repositórios. Ou seja, além de revisão de literatura, deve-se fazer uma revisão de dados existentes, relacionando-os ao contexto da pesquisa em execução. Na comunicação de resultados, por sua vez, pesquisadores devem depositar os seus dados da pesquisa em repositórios como forma de promover a reprodutibilidade e o reaproveitamento em novos estudos.

Esse modelo se apresenta alinhado a práticas ligadas aos dados abertos, na medida em que ocorre nos processos de pesquisa e na disseminação dos resultados, em um processo cíclico, no qual dados existentes apoiam a criação de novos. Murray-Rust (2008), discutindo os dados abertos na ciência, ressalta que eles têm como princípio o compartilhamento para o reúso, de forma a possibilitar novas percepções, reagrupamentos, adições etc. Assim, o movimento Dados Abertos representa a remoção de barreiras no compartilhamento de dados de pesquisas, como em uma evolução do Acesso Aberto, que removeu as barreiras para o acesso aos artigos resultados de pesquisas.

Pampel e Dallmeier-Tiessen (2014) advogam pelo que chamam de nova ciência, feita pelo compartilhamento e pela diminuição de barreiras, revelando a necessidade de criação de estratégias para o fomento e a abertura dos dados frente à crescente demanda da comunidade por tais práticas. Para os referidos autores, um dos pontos primordiais para o sucesso dos dados abertos é a gestão dos dados de pesquisa por meio de infraestrutura ancorada em políticas robustas que apoiem a cooperação entre estudiosos, na medida em que mudanças nas práticas científicas só ocorrem com a aceitação da comunidade acadêmica.

Evidentemente, os desafios não remetem apenas ao comportamento dos pesquisadores. Borgerud e Borglunf (2020), por exemplo, relatam que na Suécia, mesmo com regulamentação federal, a abertura dos dados de pesquisa enfrenta desafios, como a questão de preservação por longos períodos de tempo, numa visão Mertoniana do arquivamento dos dados, baseados nos pilares da acessibilidade,

preservação, verificação e reusabilidade. Esse estudo, apesar de preliminar, reverbera o que muitos países enfrentam, refletindo que não basta à comunidade aceitar uma nova prática, somada à atuação governamental na implementação de leis e outras regulações de apoio; fazem-se necessários também estudos que amparem a criação de infraestrutura que atendam e amparem essas novas práticas.

Outro ponto discutido na abertura dos dados de pesquisa é a qualidade. Nessa frente, Koltay (2020) discute a confiabilidade dos dados com várias relações e critérios que ajudam na sua verificação, como: originalidade, métodos de coleta e processamento, autenticidade, aceitabilidade, aplicabilidade, compreensibilidade, entre outros. Da mesma forma, ressalta preocupações com qualidade técnica em relação às bases de dados compartilhadas, com vistas a possibilitar o reúso dos dados nelas dispostos. Nesse ponto, reitera a necessidade de curadoria dos dados para garantir integridade e autenticidade, a fim de permitir que pesquisadores possam acessar bases de dados confiáveis e seguras.

Nesse contexto, revela-se que os dados abertos possuem complexidades e desafios na sua implementação, envolvendo questões culturais, procedimentais, tecnológicas, legais, entre outras. A eliminação de barreiras para o compartilhamento dos dados de pesquisa requer estudos em várias áreas, os quais devem criar uma vasta base conceitual que orientará o enfrentamento dos diversos desafios. Com isso, deve-se assegurar que as ações, métodos e tecnologias procurem atender, de forma eficaz e satisfatória, a finalidade de reúso desses dados.

3. Dados pesquisa

O conceito dados de pesquisa, conforme apontou Costa (2017), pode estar afeito ou influenciado por variadas categorias que são atribuídas aos próprios dados, ocasionando, assim, o surgimento de variadas definições que espelham a tipologia dos dados, suas formas de geração, coleta ou acesso, suas finalidades, estágios de pesquisa em que os dados são utilizados ou gerados etc.

Nesse sentido, dados de pesquisa diferem conforme a disciplina, em grande parte por causa dos métodos utilizados, na medida em que geralmente as ciências rígidas possuem certa predileção pelos métodos quantitativos, ao passo que as humanidades, por sua vez, preferem métodos mais qualitativos. Assim, os dados de pesquisa refletem os procedimentos metodológicos, gerando dados mais ou menos estruturados e compondo bases de dados com grandes diferenças, que refletem a natureza da disciplina, criando desafios diferenciados.

Logo, as disciplinas devem possuir recomendações diferenciadas, dependendo do tipo de base de dados. Especificamente nas ciências sociais, Diaz (2019) recomenda que o compartilhamento de dados requer preocupações com a regulamen-

tação vigente do país, com o código de ética e leis que orientam desde a coleta até a publicação, assegurando a permissão do pesquisador no compartilhamento dos dados, verificações dos riscos, apoio da sua instituição no processo, e ter ciência das responsabilidades éticas da pesquisa e dos dados coletados.

Nessa mesma linha, Sayão e Sales (2016) observam que "o termo 'dado de pesquisa' tem uma amplitude de significados que vão se transformando de acordo com domínios científicos específicos, objetos de pesquisas, metodologias de geração e coleta de dados e muitas outras variáveis".

Uma das definições mais recorrentemente citadas na literatura é a apresentada pelo National Science Board dos Estados Unidos (2011), segundo o qual dados de pesquisa são registros factuais em meio digital, geralmente aceitos pela comunidade acadêmica para fins de validação de resultados de pesquisa. Tal definição pode ser tomada como restritiva, uma vez que exclui a possibilidade de dados como objetos físicos e, nesse caso, inclusive objetos documentais, além de toda uma outra categoria de registros materiais, não digitalizados ou computacionais, que podem ser úteis no contexto das humanidades, por exemplo. Apesar disso, a comunicação, em geral, requer a transposição ou diferentes representações para outras formas de registro, com maior carga lógica e conceitual. Nesse sentido, a definição mais ampla, apresentada por Sales e Sayão (2019), auxilia a acomodar possíveis variações de contexto. Esses autores definem dados de pesquisa como "todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito como necessário para validar os resultados da pesquisa pela comunidade científica" (sales; sayão, 2019, p. 36).

A predileção pela natureza digital dos dados não é nova, podendo remeter ao conceito de *datalogy*, originalmente apresentado por Peter Naur em 1966 para se referir ao estudo e processamento de dados em ambientes já computadorizados (naur, 2007; zhu; xiong, 2015). Mais recentemente, Zhu e Xiong (2015) também problematizaram uma distinção interessante, com vistas ao delineamento de um objeto de estudo da ciência de dados, baseado na separação entre os fenômenos naturais (*real nature*), ou observáveis a partir do mundo natural ou real dos fenômenos digitais, que os autores denominaram *data nature*, ou natureza de dados, em tradução livre.

Ainda com vistas à definição e ao delineamento do que caracteriza dados de pesquisa, Costa (2017) destacou alguns conceitos e práticas importantes nesse aspecto, entre eles a gestão e o ciclo vida de dados de pesquisa. A autora ressalta que a gestão envolve processos de planejamento, manipulação, armazenamento e preservação dos dados de pesquisa, ao passo que Sayão e Sales (2016) falam sobre ações que coletivamente permeiam o ciclo de vida dos dados de pesquisa, além

da aplicação de padrões de ampla aceitação por variadas instâncias acadêmicas ou disciplinares. Sales, Costa e Shintaku (2020) apresentam três potenciais contextos de aplicação que da gestão de dados de pesquisa, a saber: o contexto dos próprios pesquisadores, nos dados que amparam suas pesquisas; o contexto das agências de fomento, focadas no impacto e na contribuição das pesquisas, potencialmente mensurável por meio do compartilhamento dos dados; e o contexto dos editores científicos, focado na verificação e reprodutibilidade de resultados de pesquisa por meio de dados. O ciclo de vida dos dados, por sua vez, pode incluir número variado de etapas que varia conforme a complexidade do ciclo, que geralmente se inicia com a etapa de geração ou coleta, até etapas de reúso ou descarte, quando for o caso.

Considerando-se o universo das tecnologias disponíveis no contexto dos dados, faz-se importante levar em conta as diferentes formas ou mídias de registro dos dados para representação de uma determinada realidade, fatores esses que podem influenciar a capacidade ou as condições de acesso ou uso dos dados de pesquisa. Viabilizar que dados transitem entre tecnologias, mantendo seus aspectos qualitativos e suas condições de reprodutibilidade e preservação, bem como a análise e interpretação desses dados, requer uma carga do que se pode chamar de competência em dados ou um nível satisfatório da chamada alfabetização em dados (*data literacy*).

Baykoucheva (2015) apresenta essa alfabetização como habilidades para a leitura, interpretação e compreensão de dados, ressaltando a importância da incorporação de programas voltados para a alfabetização em dados e informação, com elenco de conteúdo e competências específicas. Destaca que muitos aspectos que já vêm sendo trabalhados com vistas à alfabetização de dados têm proximidade com ou englobam uma alfabetização estatística, tratando do emprego de pensamento crítico sobre estatística descritiva. Calzada Prado e Marzal (2013), por sua vez, ressaltam a emergência de treinamento em competências para além do escopo estatístico, englobando competências para tarefas de aquisição de dados, avaliação, tratamento e processamento, análise e interpretação. Nesse sentido, faz-se importante a compreensão e o domínio do universo de tecnologias que amparam tarefas dessa natureza.

4. Tecnologias e os critérios FAIR

Segundo o modelo de comunicação científica de Björk (2007), dados de pesquisa podem ser buscados e compartilhados em repositórios, dependendo do tipo e do momento em que a pesquisa está. Assim, os repositórios tornam-se os sistemas de informação mais apropriados para ofertar funcionalidades como depósito e recuperação de bases de dados. Da mesma forma, podem oferecer também funcionalidades para gestão de bases de dados de pesquisa, envolvendo outros serviços.

No entanto, repositórios originalmente eram vistos no Acesso Aberto como a Via Verde (HARNAD *et al.*, 2004), em que cópias de artigos publicados em revistas eram disponibilizados livremente, tanto que Weitzel (2006) considerava os repositórios como segunda fonte. Com a possibilidade de gerir bases de dados de pesquisa, repositórios assumem significação diferente, adicionando funcionalidades apropriadas para esse tipo de objeto digital.

A verificação de adequação do software para a gestão de base de dados pode ser feita de várias formas, com modelos de avaliação diversos, dependendo dos objetivos. No presente estudo utiliza-se o atendimento aos princípios FAIR, acrônimo de *findable, accessible, interoperable, reusable*. Para tanto, considera-se:

- **Localizável**: possibilita descrição por metadado e permite o uso identificador persistente.
- **Acessível**: pode ser acessado por humanos ou máquinas, ofertar licenças claras e disponibilizar protocolos de comunicação.
- **Interoperável**: padrão de metadados pode ser entendido por máquinas e funcionalidades para entendimento das bases de dados.
- **Reusável**: base de dados descrita para possibilitar o seu reúso.

Evidentemente, os critérios básicos FAIR possuem desdobramentos, expandindo a sua abrangência conceitual. Da mesma forma, a gestão de dados com base nestes princípios envolve determinadas atividades, como a questão de curadoria de dados, que extrapolam puramente as ferramentas em si, envolvendo atividades, métodos e padrões. Assim, repositórios são instrumentos, como indicados no modelo de comunicação científica de Bjork (2007), para depósito e recuperação de bases de dados, assumindo forma de compartilhamento.

Assim, discutindo tecnologias sob os aspectos relacionados aos critérios FAIR, cada critério, incluindo os seus refinamentos como apresentado por Hanning *et al* (2019), tornam-se requisitos que os softwares para repositórios devem atender. Com isso, possibilita-se a discussão dos critérios FAIR sob a observação da tecnologia, com vistas a propor um entendimento sobre o tema, pois a tecnologia oferta ferramentas, enquanto os critérios representam a base conceitual, que possibilita a avaliação das ferramentas.

Já o critério de Localizável (*Findable*) trata inicialmente das bases de dados e seus metadados. Assim, para atender a esse critério, os repositórios tem de fornecer suporte para depósito de bases de dados, independentemente de seu formato ou tipo, e adotar padrão de metadados que possam ser acessados por pessoas e máquinas. Grande parte das tecnologias atuais para construção de

repositórios atendem a esse requisito de forma geral, mas requerem algumas observações.

O depósito das bases de dados requer alguns aspectos de curadoria, como verificação de não conter vírus ou, mesmo, mecanismos de integridade. No entanto, nas recomendações de encontrabilidade as bases de dados têm destaque menor que em relação aos metadados, principalmente com a possibilidade de ter esses metadados indexados por motores de busca. Para os metadados, questões básicas, como uso de identificador persistente, são obrigatórias. Esse ponto é relativamente atendido por sistemas de identificadores como o *Handle* ou *Digital Object Identifier* (DOI). Entretanto, as tecnologias devem possibilitar a aplicação de metadados enriquecidos, voltados para algumas atividades de curadoria, que podem apresentar desafios para os depositantes, visto que as tecnologias são campos a mais a serem implementados. Assim, nota-se que a recomendação de localizável afeta alguns pontos relacionados à qualidade dos metadados, os quais transcendem a ferramenta.

Essa visão, na qual a localização tem aspectos com impacto menor na tecnologia, pode ser amparada pelo estudo de Monteiro e Sant'Ana (2020), que apresenta soluções tecnológicas para atender aos critérios FAIR de acessibilidade. Para os autores, os princípios FAIR foram implementados na infraestrutura de pesquisa CLARÍN, incluindo aspectos de localizável por meio de uso de padrões atualmente implementados em várias tecnologias para criação de repositórios, como uso de padrões de metadados Dublin Core, que possuem flexibilidade contém padrões mínimos para interoperabilidade, segundo o movimento dos arquivos abertos (*Open Archives Initiative* - OAI).

De forma geral, para atender o princípio FAIR de localizável, tecnologias para repositórios devem permitir a indexação por motores de buscas, implementar padrões de metadados flexíveis e interoperáveis, que possibilitem o uso de identificadores persistentes para as bases de dados, de forma a ser encontrado facilmente. Assim, caso não haja revisões nesses critérios, grande parte das tecnologias de repositórios os atendem, em maior ou menor grau, visto que repositórios, na sua função tradicional em comunicação científica, têm a função de facilitar o acesso aos seus itens.

Já para o critério acessível, os aspectos tecnológicos são maiores, visto terem relação com os acessos às bases de dados por meio de protocolos de comunicação ou diretamente para o reúso. Bases de dados depositadas em repositórios estão em formato digital dependentes do seu tipo e, mesmo que podendo ser consideradas dados brutos, podem apresentar inúmeros formatos. Assim, os repositórios podem disponibilizar os arquivos para serem baixados (*download*) ou visualizados

por *streaming*, dependendo das permissões. No caso da visualização, repositórios precisam ofertar funcionalidades optativas ante aos inúmeros formatos que bases de dados de pesquisa podem assumir, como tabelas, planilhas, áudios, vídeos, texto e outros.

O acesso aos recursos se dá por meio de oferta de acesso, possibilitando a interação, com uso de protocolos de comunicação. Esses critérios têm relação com a infraestrutura na qual os repositórios serão hospedados. Esse critério está associado ao uso de ferramentas livres para criação de ambientes, como o uso de sistemas operacionais, servidores de aplicação e banco de dados livres. Não é apenas usar o software livre para criar o repositório, mas ter todo o ambiente construído com tecnologias livres.

O único ponto relacionado à tecnologia para criação de repositório e o critério de acesso tem relação com a preservação dos metadados, mesmo que a base de dados não esteja mais disponível. Esse ponto deve orientar a remoção de bases de dados de repositórios por qualquer motivo, obrigando a manter os metadados. Assim, pode ter impacto nos metadados, requerendo a presença de campo indicador do status da base de dados.

Os critérios relacionados à interoperabilidade são totalmente tecnológicos. Desde o início do movimento de arquivos abertos, ainda no final do século passado, muitos softwares vêm implementando o protocolo *Open Archives Initiative* (OAI) nas suas versões *Protocol Metadata Harvesting* (PMH) ou *Object Reuse and Exchange* (ORE), baseadas em *eXtensible Markup Language* (XML). Entretanto, com a evolução tecnológica, outras formas de notação podem ser utilizadas para a interoperabilidade, como a tecnologia *JavaScript Object Notation* (JSON), ou mesmo com o uso de serviços *WebServices*, com bons resultados e maior flexibilidade na coleta de metadados.

Nesse sentido, a ideia da interoperabilidade é maior nos critérios FAIR que nos arquivos abertos, voltados à possibilidade de intercâmbio de metadados e objetos digitais. Assim, tecnologias para implementação de repositório de dados de pesquisa precisam ser ajustadas para atender ao FAIR. Mais que isso, requer desenvolvimento de estruturas de apoio como ontologias e tesauros, padronizando conteúdo de campos de metadados com vocabulários controlados. Da mesma forma, irá requerer dos produtores de dados que depositem os dados, dicionário de dados e, futuramente, narrativas de dados.

Mesmo que os critérios de interoperabilidade tenham aspectos tecnológicos, nota-se a presença de recomendações fortes voltadas a metadados, como nas questões de relacionamento entre bases de dados. Assim, se uma base de dados é interpelada, a fonte original pode ser citada indicando a sua proveniência. Da mesma

forma, se uma base de dados é gerada com base em outras, isso deve ser indicado. Todas essas informações devem estar previstas nos metadados.

Para que bases de dados sejam reusáveis são necessárias descrições completas, que extrapolam os metadados. Por isso, os critérios FAIR para reúso orientam a desenvolver narrativas de dados que compreendem limitações, contexto, formas de coleta, entre outros. As recomendações para reúso têm aspectos procedurais, em que o exposto são informações que permitirão o seu reúso.

Cabe assim destacar que os critérios FAIR orientam produtores de dados que desejam compartilhar e aderir aos movimentos de dados abertos e ciência aberta, pesquisadores que promovam a transparência pela disseminação dos dados de pesquisa. Nesse sentido, os critérios FAIR se apresentam como uma base de orientação geral para todas as iniciativas que desejam adotar o compartilhamento de dados como parte das atividades de pesquisa.

5. Tecnologias para repositório FAIR

Se os critérios FAIR têm aspectos de orientação, o documento FAIR *Data Point Specification (*FDPS*)* apresenta os requisitos para desenvolvimento ou avaliação de tecnologias voltadas ao atendimento dos princípios FAIR. Assim, seguindo as especificações descritas no referido documento, será criado um repositório que atenderá aos metadados e dados que podem ser classificados como FAIR. Da mesma forma, o documento pode ser utilizado para avaliação de conjuntos de software existentes, principalmente os livres e de código aberto, cujo objetivo é propor alterações ou ajustes.

O repositório pensado de acordo com o modelo FDPS é, em geral, voltado às ciências naturais, algumas vezes focado na biologia, o que apresenta certa restrição. Entretanto, com adaptações, pode ser utilizado para outras ciências, tendo em vista as especificidades de cada área. A intenção é promover tecnologias que possibilitem a criação de repositórios que possam interoperar metadados e bases de dados a fim de possibilitar uma pesquisa distribuída.

Assim, o repositório pensado pelo FDPD tem dois conjuntos de funcionalidades bem estabelecidos, voltados à recuperação e ao depósito de bases de dados. Esses pontos alinham-se aos repositórios atuais, voltados à disseminação de documentação técnica e científica. Nesse sentido, conceitualmente, o repositório de dados FDPD apresenta aspectos semelhantes aos repositórios digitais, só que especializado em gerir bases de dados.

Para a recuperação das bases de dados, há processos simples, como encontrar e acessar as bases de dados. O depósito de bases de dados, por sua vez, pode ser manual ou automático. Independentemente dos processos, se para recuperação ou

depósito, requer-se que o repositório ofereça funcionalidades simples e padronizadas, para que os usuários usem a plataforma para uso ou oferta de bases de dados.

Os requisitos iniciais para atender ao FDPS têm início com as funcionalidades de encontrabilidade das bases de dados, ou seja, funcionalidades de descoberta, que ajudam os usuários a descobrir onde essas bases de dados estão, envolvendo questões relacionadas à indexação por parte de motores de busca e outros. De modo semelhante, os repositórios devem oferecer ferramentas de busca simples, mas com indexação composta de funcionalidades que apoiem a descoberta dos conjuntos de dados mantidos por ele.

Uma vez encontrada a base de dados desejada, inicia-se o processo de acesso. Assim, as licenças para acesso e uso dos dados devem ser apresentadas, preferencialmente aquelas indicadas pelo FAIR. Da mesma forma, o repositório deve orientar na padronização do formato utilizado para os dados, com forte indicação para uso de formatos livres, para que possibilite acesso padronizado. Assim, usuários que desejem agregar bases de dados de repositórios diferentes têm seu trabalho facilitado, promovendo o reúso.

O depósito de dados, também chamado de publicação de dados, consiste num conjunto de funcionalidades que auxiliam os autores a disponibilizar conjuntos de dados por meio do repositório, com uma grande variedade de opções de fornecimento de acesso, versionamento, status, entre outros. Logo, visa atender às especificidades apresentadas pelas bases de dados em relação à restrição de uso, momento da pesquisa e outros, mas possibilitando que usuários descubram essas bases.

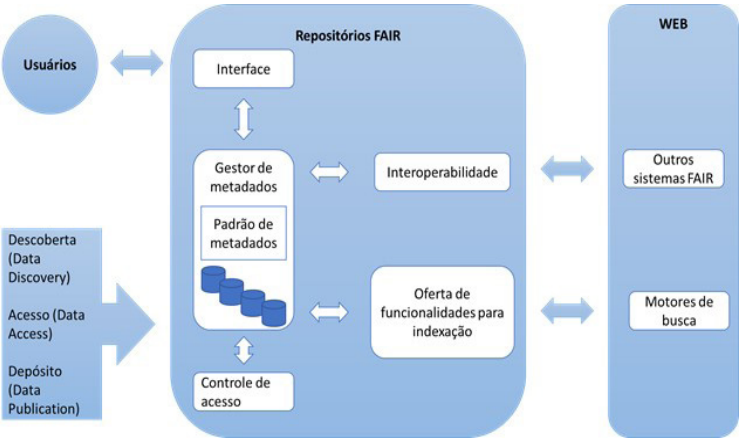
Por fim, repositórios de dados FDPS devem possibilitar a geração de estatísticas de uso. Indicadores devem ser ofertados como informação estratégica para tomada de decisão relacionada à gestão de infraestrutura, por exemplo. De modo parecido, possibilita-se conhecer a relevância das bases de dados, gerando informações para os autores sobre o reúso de seus dados. Assim, estatísticas são úteis até para justificar o repositório e os seus investimentos.

Por seu lado, como todos os repositórios, os sistemas que se alinharem ao FDP devem possuir uma arquitetura que possibilite gerenciar itens compostos pelas bases de dados e seus metadados, os quais podem ser manipulados tanto por humanos quanto por máquinas. Assim, deve possuir uma interface Web na qual usuários podem buscar, depositar e recuperar bases de dados por meio dos metadados das bases. Da mesma forma, deve fornecer acesso para que máquinas possam recuperar informações. O acesso deve ser controlado, por permissões, com vistas a proteger dados sensíveis. Os metadados devem seguir padrões FAIR, ofertando informações sobre cinco níveis: sobre o próprio repositório; sobre o catálogo (conjunto de bases

de dados que compõem o acervo); sobre a base de dados; sobre a distribuição; e sobre os registros de dados.

De modo simplificado, a arquitetura de um repositório que atenda aos princípios FAIR (Figura 1) atende a usuários humanos por meio de uma interface Web, oferecendo interoperabilidade para outros sistemas e permitindo que motores de busca indexem os seus metadados. O acesso aos metadados e as bases de dados deve ser controlado a fim de possibilitar disseminação escalonada devido a diferentes níveis de sensibilidade dos dados. Os metadados seguem padrões que facilitam a integração, descoberta e descrição das bases de dados. Em suma, o repositório FDPS facilita a descoberta, o acesso e o depósito de dados.

Figura 1 – Arquitetura de um repositório que atenda os princípios FAIR



Fonte: Elaboração dos autores baseados no FAIR *Data Point Specification* (FDPS)

As especificações para repositório FDPS não diferem muito das tecnologias utilizadas para a criação de repositórios acadêmicos, como sistemas informatizados. Entretanto, bases de dados possuem especificidades que requerem tratamento diferenciado. Como consequência, as diferenças entre bases de dados e publicações acadêmicas irão distinguir os dos sistemas, assim como os tipos de funcionalidades propostas.

6. Considerações finais

As orientações FAIR possuem aspectos que mudam a forma de analisar as práticas da ciência, na medida em que alteram o foco dos resultados de pesquisa, representados por artigos, livros, anais de eventos e outros, para os dados de pesquisa, com o objetivo de armazená-los para o seu compartilhamento. Muda, em certo

ponto, a perspectiva dos dados de pesquisa, como um bem social que deve ser compartilhado como forma de contribuir com outros cientistas.

Para tanto, requer infraestrutura informacional, composta por sistemas informatizados, atendidos atualmente por repositórios de dados de pesquisa. Assim, para implementar esses sistemas, há oferta de tecnologias livres e proprietárias, que atendem a parte dos critérios. Entretanto, a tarefa de implementar princípios FAIR na disseminação de dados de pesquisas está, até certo ponto, a cargo dos gestores e pesquisadores, visto que as tecnologias voltadas à criação de repositórios acadêmicos atendem a grande parte das demandas.

Neste sentido, reforça-se que a mudança para o uso dos princípios FAIR é comportamental, por um lado, na medida em que depende dos pesquisadores depositarem os dados dos seus estudos em repositórios de dados com licenças abertas, com metadados bem descritos e com uso de formatos livres. Depende, da mesma forma, que haja repositórios de dados bem estruturados, implementadores dos princípios FAIR, ofertados por instituições reconhecidas no mundo acadêmico, garantindo a curadoria necessária.

Para autores que adotem os princípios FAIR, pode-se inserir no corpo do artigo o link para os dados residentes em repositórios na seção de metodologia, por exemplo. Outra opção válida pode estar na revista adotar um campo de metadados para inserção deste link no processo de submissão, como muitas revistas já o fazem. Com relação à tecnologia, softwares para criação de repositórios de dados, revistas ou mesmo formatos para formatação de artigos já atendem aos princípios.

Pode ser que futuramente não se atendam a todos os princípios FAIR, pelo fato de os repositórios ainda não atenderem a todos os desdobramentos dos princípios. Grande parte dos repositórios fundamenta-se no processo de compartilhamento e não na curadoria do seu acervo. Por isso, os desafios em relação à infraestrutura tecnológica ainda requerem estudos.

Possivelmente, o maior desafio da gestão dos dados de pesquisa quanto a tecnologias relacione-se às que atendam às necessidades de curadoria, visto que a disseminação é apenas a parte final do processo, na qual o FAIR se contextualiza. Possivelmente o FAIR, em futuro próximo, se adapte às novas exigências, ao adotar critérios atendidos por tecnologias. Com isso, novas necessidades surgirão, numa contínua evolução.

7. Referências

BAYKOUCHEVA, Svetla. **Managing Scientific Information and Research Data.** Burlington: Elsevier Science, 2015.

BJÖRK, Bo-Christer. A model of scientific communication as a global distributed

information system. **Information Research**, v. 12, n. 2, p. 307, 2007. .

BORGERUD, Charlotte; BORGLUND, Erik. Open research data, an archival challenge? **Archival Science**, 10 fev. 2020. DOI 10.1007/s10502-020-09330-3. Disponível em: http://link.springer.com/10.1007/s10502-020-09330-3. Acesso em: 3 nov. 2020.

CALZADA PRADO, Javier; MARZAL, Miguel Ángel. Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. **Libri**, v. 63, n. 2, jan. 2013. DOI 10.1515/libri-2013-0010. Disponível em: https://www.degruyter.com/doi/10.1515/libri-2013-0010. Acesso em: 3 nov. 2020.

COSTA, Michelli Pereira da. **Fatores que influenciam a comunicação de dados de pesquisa sobre o vírus da zika, na perspectiva de pesquisadores**. 2017. 269 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Faculdade de Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação, Brasília, 2017. Disponível em: https://repositorio.unb.br/handle/10482/23000. Acesso em: 9 set. 2020.

DIAZ, Pablo. Ethics in the era of open research data: some points of reference. FORS **Guide**, 2019. DOI 10.24449/FG-2019-00003. Disponível em: https://forscenter.ch/fors-guides/fg-2019-00003/. Acesso em: 3 nov. 2020.

HARNAD, Stevan; BRODY, Tim; VALLIÈRES, François; CARR, Les; HITCHCOCK, Steve; GINGRAS, Yves; OPPENHEIM, Charles; STAMERJOHANNS, Heinrich; HILF, Eberhard R. The Access/Impact Problem and the Green and Gold Roads to Open Access. **Serials Review**, v. 30, n. 4, p. 310–314, jan. 2004. DOI 10.1080/00987913.2004.10764930. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/00987913.2004.10764930. Acesso em: 3 nov. 2020.

HENNING, Patricia Corrêa; RIBEIRO, Claudio José Silva; SANTOS, Luiz Olavo Bonino da Silva; SANTOS, Paula Xavier Dos. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389–412, 26 abr. 2019. DOI 10.19132/1808-5245252.389-412. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/84753. Acesso em: 9 set. 2020.

KOLTAY, Tibor. Quality of Open Research Data: Values, Convergences and Governance. **Information**, v. 11, n. 4, p. 175, 25 mar. 2020. DOI 10.3390/info11040175. Disponível em: https://www.mdpi.com/2078-2489/11/4/175. Acesso em: 3 nov. 2020.

MONTEIRO, Elizabete Cristina de Souza de Aguiar; SANT'ANA, Ricardo Cesar Gonçalves. Repositórios de Dados Científicos na Infraestrutura de Pesquisa:

adoção dos princípios FAIR. **Ciência da Informação**, v. 48, n. 3, mar. 2020. Disponível em: http://revista.ibict.br/ciinf/article/view/4878. Acesso em: 3 nov. 2020.

MURRAY-RUST, Peter. Open Data in Science. **Nature Precedings**, 18 jan. 2008. DOI 10.1038/npre.2008.1526.1. Disponível em: http://www.nature.com/articles/npre.2008.1526.1. Acesso em: 3 nov. 2020.

NATIONAL SCIENCE BOARD. **Digital Research Data Sharing and Management**. Arlington, Virginia: National Science Foundation, 2011. Disponível em: https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf. Acesso em: 3 nov. 2020.

NAUR, Peter. Computing versus human thinking. **Communications of the ACM**, v. 50, n. 1, p. 85–94, jan. 2007. DOI 10.1145/1188913.1188922. Disponível em: https://dl.acm.org/doi/10.1145/1188913.1188922. Acesso em: 3 nov. 2020.

PAMPEL, Heinz; DALLMEIER-TIESSEN, Sünje. Open Research Data: From Vision to Practice. *In*: BARTLING, Sönke; FRIESIKE, Sascha (orgs.). **Opening Science**. Cham: Springer International Publishing, 2014. p. 213–224. DOI 10.1007/978-3-319-00026-8_14. Disponível em: http://link.springer.com/10.1007/978-3-319-00026-8_14. Acesso em: 3 nov. 2020.

SALES, Luana Farias; COSTA, Michelli; SHINTAKU, Milton. Ciência aberta, gestão de dados de pesquisa e novas possibilidades para a editoração científica. *In*: SHINTAKU, Milton; SALES, Luana Farias; COSTA, Michelli (orgs.). **Tópicos sobre dados abertos para editores científicos**. 1. ed. Botucatu, SP: ABEC, 2020. p. 13–21. DOI 10.21452/978-85-93910-04-3.cap1. Disponível em: https://www.abecbrasil.org.br/arquivos/Topicos_dados_abertos_editores_cientificos.pdf#01. Acesso em: 3 nov. 2020.

SALES, Luana Farias; SAYÃO, Luís Fernando. Uma proposta de taxonomia para dados de pesquisa. **Revista Conhecimento em Ação**, v. 4, n. 1, p. 31–48, 30 jun. 2019. DOI 10.47681/rca.v4i1.26337. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337. Acesso em: 3 nov. 2020.

SAYÃO, Luis Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, v. 21, n. 2, p. 90, 2016. DOI 10.5433/1981-8920.2016v21n2p90. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 9 set. 2020.

SOMPEL, Herbert Van de; LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. **D-Lib Magazine**, v. 6, n. 2, 2000. DOI 10.1045/february2000-vandesompel-oai. Disponível em: http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html. Acesso em: 11 mar. 2020.

SULEMAN, Hussein; FOX, Edward A. A Framework for Building Open Digital Libraries. **D-Lib Magazine**, v. 7, n. 12, dez. 2001. DOI 10.1045/december2001-suleman. Disponível em: http://www.dlib.org/dlib/december01/suleman/12suleman.html. Acesso em: 3 nov. 2020.

WEITZEL, Simone da Rocha. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em questão**, v. 12, n. 1, p. 51–71, 2006. .

ZHU, Yangyong; XIONG, Yun. Towards Data Science. **Data Science Journal**, v. 14, n. 0, p. 8, 22 maio 2015. DOI 10.5334/dsj-2015-008. Disponível em: http://datascience.codata.org/article/10.5334/dsj-2015-008/. Acesso em: 3 nov. 2020.

Como citar com o DOI individual

SHINTAKU, Milton; APPEL, André Luiz; OLIVEIRA, Alexandre Faria de. Tecnologias para gestão de dados de pesquisa segundo preceitos FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 129 - 146. DOI: 10.22477/9786589167242.cap10

Seção 3

DADOS INTEROPERÁVEIS

Interoperabilidade de dados e a transdução informacional encapsulada no acesso a dados

Ricardo César Gonçalves Sant'Ana[1]

1. Introdução

Diminuir a distância e as barreiras entre os dados e as necessidades informacionais é o desafio que se apresenta. E quando se considera o cenário de acesso a dados, esta questão inclui, também, a necessidade de localização, interpretação e uso não somente por indivíduos mas principalmente por máquinas.

Esta demanda implica tornar estes dados em conteúdos interpretáveis (legíveis), em situações e momentos distintos daqueles em que foram criados, gerando a necessidade de adoção de padrões e, por consequência, princípios e diretrizes que, compartilhados, propiciem reuso de acervos de dados.

Ambientes de alta demanda por acesso a dados, tais como os corporativos ou mesmo relacionados à gestão pública, são compostos por centenas ou mesmo milhares de sistemas, desenvolvidos interna ou externamente, compondo um cenário complexo e diversificado, todos requerendo integração de seus dados para que seu efetivo uso seja viabilizado (reeve, 2013, sant'ana, 2009; dyché & levy, 2006, p.384). No entanto, grande parte do foco da gestão dos dados é aplicada nos processos de coleta, armazenamento e disponibilização nos sistemas em detrimento ao fluxo de dados entre as diferentes estruturas (reeve,2013). Esse espaço entre os sistemas tende a aumentar exponencialmente sua complexidade com o aumento de fontes de dados consideradas no ambiente e é geralmente atendido por interfaces de dados (recursos desenvolvidos para viabilizar o fluxo de dados entre sistemas).

Assim, ganha relevância crescente a busca pela disponibilização dos dados no lugar certo, no momento certo, no formato certo, e aderentes à necessidade informacional que se pretende atender, o que coloca a integração dos dados como condição central ao sucesso dos processos de acesso a dados (kelleher & tier-

1 Livre-Docente em Sistemas de Informações Gerenciais, Professor Adjunto na Universidade Estadual Paulista - unesp, ricardo.santana@unesp.br

ney, 2018, p.346). Essa integração dos dados implica, por sua vez, em processos de transformação destes dados, de forma cada vez mais automatizada, para que esses dados possam ser tratados como um conjunto único, gerando a necessidade de transduções informacionais (sant'ana, 2019), que por sua vez, exigem definições que derivam de conhecimentos tanto técnicos quanto sobre o contexto (reeve, 2013; shkedi,2019, p.32). Na dimensão do conhecimento sobre o contexto, emergem linhas de enfrentamento aos desafios da integração, como as que consideram, por exemplo, o uso de ontologias, mas esta questão foje ao escopo deste texto.

Considerando o cenário de múltiplos sistemas, para múltiplas instâncias, tal como quando se pensa em compartilhar dados entre academia, indústria, agências de financiamento e editoras acadêmicas, pode-se prever o alto grau de complexidade que se apresenta. Buscando o desenvolvimento de meios compartilhados de se ampliar o reuso de acervos de dados para situações como esta, representantes destas instâncias se reuniram em um workshop realizado em Leiden, Holanda, em 2014, denominado 'Joint Designing a Data Fairport' para "projetar e endossar, em comum acordo, um conjunto conciso e mensurável de princípios que chamamos de Princípios de Dados fair." (wilkinson et al, 2016), denominação resultante dos conceitos: *Findable, Accessible, Interoperable e Reusable* (Encontrável, Acessível, Interoperável e Reutilizável). Um diferencial proposto pelas diretrizes dos Princípios fair está no fato de que "enfatizam especificamente o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente, além de apoiar sua reutilização por indivíduos" e também cabe destaque a intenção de que as diretrizes sejam aplicáveis também aos "algoritmos, ferramentas e fluxos de trabalho que levaram a esses dados" (wilkinson et al, 2016).

Os Princípios fair, após ajustes e aprimoramentos, apresentam-se compostos por quatro princípios, cada um deles com critérios próprios: quatro relacionados ao conceito *Findable,* dois ao *Accessible*, um ao *Reusable* e três ao *Interoperable.* Estes últimos, foco deste capítulo, sendo eles (wilkinson et al, 2016):

I1. metadados com uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento;
I2. metadados com vocabulários que seguem os princípios fair;
I3. metadados incluem referências qualificadas a outros metadados.

São princípios genéricos mas que apontam para diretrizes que podem contribuir, e muito, na ampliação do potencial de interoperabilidade dos conjuntos de dados. Aspectos como a necessidade de formalismo, ou seja, de atendimentos a padrões pré estabelecidos, pode ajudar a minimizar definições divergentes do mesmo

conteúdo. O conceito 'acessível', ainda previsto no princípio I1, também pode corroborar com a necessidade de que estes padrões e regras sejam compartilhados entre os envolvidos e que, ainda, sejam alvo de estratégias de difusão para sua compreensão, elementos que levam ao conceito, também parte do I1, de 'compartilhada'. Todos estes fatores poderiam não ter sua relevância se não fosse considerada a viabilidade de tais definições, o que leva ao conceito de 'aplicável' que completa o princípio I1.

O princípio I2 está focado na questão dos vocabulários utilizados nos metadados definidos para os conjuntos de dados, apontando de forma recursiva aos demais princípios FAIR. Esta questão é bastante ampla e requer lembrar que os vocabulários nem sempre podem atender às necessidades específicas, o que pode levar a publicações de extensões dos vocabulários existentes ou mesmo a criação de novos vocabulários (FORCE11, 2020), o que não deixa de ser um fator complicador.

Já o princípio I3 indica a necessidade de referências qualificadas entre os metadados (WILKINSON et al, 2016) o que aponta para a necessidade de que recursos maquínicos sejam capazes de realizar operações diretamente sobre os dados coletados, o que por sua vez, requer que os metadados "devem ser sintaticamente analisáveis e semanticamente acessíveis por máquina" (FORCE11, 2020). A FORCE11 aponta, ainda, que a sintaxe e a 'semântica' dos modelos e formatos de dados usados para metadados devem ser "fáceis de identificar e usar, analisar ou traduzir por máquinas", e este é um dos elementos norteadores para as argumentações apresentadas neste texto.

Nesta mesma linha de análise, destaca-se uma flexibilização prevista na proposta dos Princípios FAIR por meio de uma possibilidade de definições emergirem em um movimento *bottom-up,*

> "se um provedor pode provar que um modelo / formato alternativo de dados é inequivocamente analisável para um dos formatos FAIR adotados pela comunidade, não há nenhuma razão particular para que tal formato não possa ser considerado FAIR. Alguns tipos de dados podem simplesmente não ser 'capturáveis' em um dos formatos existentes e, nesse caso, talvez apenas parte dos elementos de dados possam ser analisados" (FORCE11, 2020)

Tal flexibilização ampliaria o potencial de aderência das especificidades inerentes ao grande número de situações e contextos de origem dos dados, ao mesmo tempo que carrega consigo a complexidade da qual se originou a motivação para a própria proposta dos Princípios FAIR. Este efeito é previsto, inclusive, pela própria FORCE11 ao propor: "a situação ideal é restringir a publicação de dados FAIR ao mínimo possível de formatos e padrões adotados pela comunidade", considerando

que seria necessário oferecer soluções às novas demandas: "A FAIRports oferecerá cada vez mais orientação e assistência nesses casos."(FORCE11, 2020).

Mesmo considerando que não seja pré-requisito para determinação de aderência dos dados aos Princípios FAIR (WILKINSON et al, 2016), deve-se buscar o acesso maquínico com maior a autonomia possível - acesso e interpretação a tal ponto que seja possível transformar os dados coletados em um novo conjunto de dados mais aderente à cada necessidade (REEVE, 2013). Não há como não considerar níveis mínimos de autonomia de processamento maquínico como condição sem a qual o processo de acesso a dados não teria como fazer frente a oferta de dados a qual estamos submetidos, ou como previsto pela FORCE11 (2020), quando afirma que "fornecer dados legíveis por máquina como o substrato principal para a descoberta do conhecimento [...] que funcionem sem problemas e de forma sustentável é um dos grandes desafios da eScience".

Mas de onde se origina tamanha variedade de elementos informacionais necessários para que os dados possam ser devidamente coletados e transformados para o uso? Em grande parte de sua própria essência fragmentada, resultante da necessária estrutura, nativa ou obtida após tratamento, mas sempre requisito para que os algoritmos possam estabelecer de forma unívoca e detalhada, passo a passo, o que a máquina deve fazer no processamento dos conteúdos.

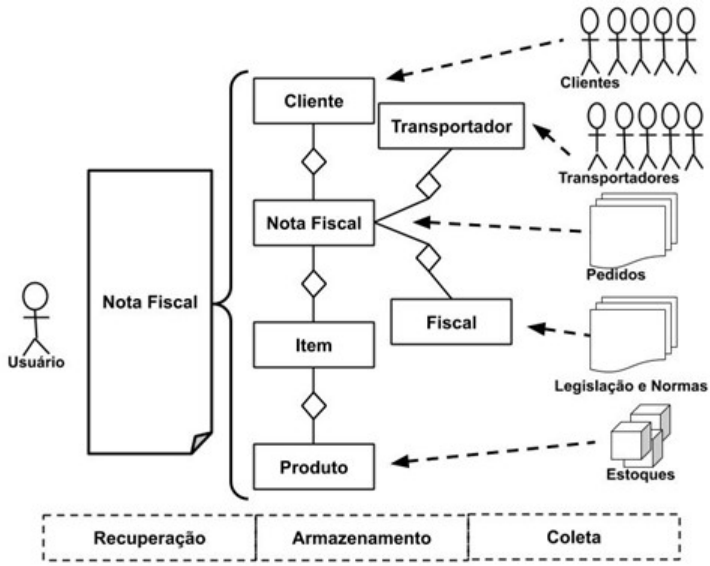
2. A natureza fragmentada dos dados e os princípios FAIR

Por sua natureza, os recursos maquínicos são capazes apenas de executar instruções absolutamente detalhadas e precisas, de onde deriva a necessidade de se estabelecer, exata e formalmente, o que deve ser feito com cada partícula informacional dos conteúdos a serem tratados. Essa algoritmização do processo leva a uma necessária e inevitável natureza fragmentada dos dados, de forma que se possa estabelecer novas camadas informacionais com metadados que sustentam elementos sintáticos e semânticos a cada um dos fragmentos, compondo assim, uma estrutura mínima de significação - a tríade: entidade, atributo e valor <e,a,v> (SANTOS e SANT'ANA, 2015) - que propicia, por sua vez, a viabilidade da algoritmização do tratamento maquínico dos conteúdos.

Esta fragmentação gera, já na fase de coleta (SANT'ANA, 2016), a necessidade de alocação dos valores específicos, relativos a cada uma das transações e fatos registrados, em 'atributos' específicos, que por sua vez estarão vinculados à 'entidades' relacionadas a cada uma das informações identificadas como relevantes. Essa vinculação emerge de mapeamento lógico dessas entidades, respeitados princípios como os relacionados à normalização dos dados, evitando redundâncias e trazendo coerência para os conjuntos de dados.

Assim, se considerarmos, como exemplo, um documento referente a uma transação de venda, uma Nota Fiscal (Figura 1), teremos toda uma trajetória dos dados, partindo da obtenção, na fase de coleta (SANT'ANA, 2016), com a identificação de informações sobre o cliente, detalhes do pedido, dos produtos e quantidades envolvidas, transportador responsável pela entrega, classificações e cálculos fiscais, entre outras informações. Estas informações são então registradas, persistidas, nas respectivas estruturas semânticas (entidades) com seus respectivos rótulos (atributos) e, ainda, relacionadas entre si, de tal forma que já passa a ser possível sua visualização como 'um' documento.

Figura 1 - Estrutura Fragmentada dos dados



Fonte: Elaborado pelo autor.

A partir deste documento, que passa a ser disponibilizado na fase de recuperação, o usuário passa a ter a possibilidade de visualizar os dados da transação, portanto, do fato. É, também, esse resultado da composição dos respectivos dados de cada entidade (conjunto de dados) e disponibilizado como um documento, que será compartilhado com os demais envolvidos, tais como o próprio cliente, o fisco, o transportador, e ainda, os demais sistemas da própria organização, tais como o sistema financeiro, o sistema contábil, estoque, entre outros.

Desta fragmentação emergem dois pontos de reflexão apontados neste texto: a transdução informacional e o encapsulamento da complexidade.

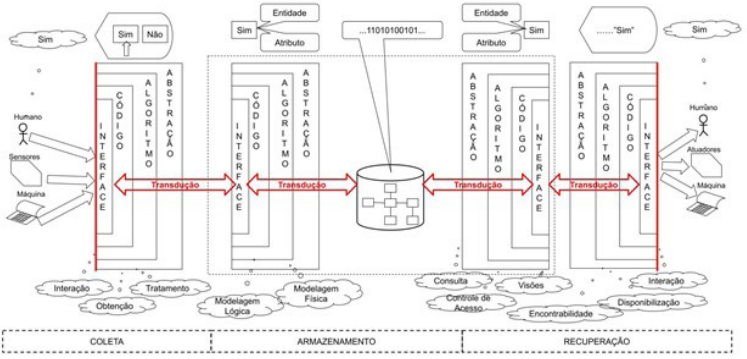
Os conteúdos, em fluxo entre fonte e usuário, coletados, armazenados e disponíveis à recuperação em distintos ciclos de vida dos dados (sant'ana, 2016), sofrem transformações, tanto no âmbito da energia quanto do formato e até do próprio conteúdo. Essas transduções informacionais permitem que a necessidade informacional do usuário seja atendida, respondendo à demandas por maior aderência com o contexto de uso, tais como: personalização, adequação, adaptação, resultando no menor custo possível ao acesso a dados.

Mesmo sendo de pleno domínio por aqueles que atuam nas camadas de abstração mais próximas ao maquínico - analistas, desenvolvedores e administradores, seja na dimensão de programação (*software*), seja na dimensão de dados - a maior parte dos usuários não têm a percepção desta estrutura fragmentária a qual os dados são submetidos para que possam ser utilizados, e nem poderiam ter, já que tal complexidade inviabilizaria o uso dos recursos computacionais.

2.1. Transdução Informacional

O trajeto dos dados é longo e complexo, partindo de sua apreensão a partir do fato, transitando pelas diversas transformações impostas pelas interfaces e adequações aos modelos de estrutura de dados, chegando ao seu registro nos suportes digitais para, finalmente, serem recompostos em um formato de documento (como no exemplo apresentado na figura 1). Cada uma destas transformações (Figura 2) implica conversões não só de forma e conteúdo como, inclusive, de energia. Tais transformações, aqui definidas como Transduções Informacionais (sant'ana, 2019) são necessárias para que os mecanismos mediadores possam tratar os conteúdos, no entanto, ampliam a complexidade do processo, tornando sua compreensão bastante custosa, o que inviabilizaria a utilização dos sistemas envolvidos.

Figura 2 - Transdução Informacional no Acesso a Dados



Fonte: Elaborado pelo autor.

Essa eventual barreira é superada pelo ocultamento de detalhes não essenciais aos utilizadores, origem do segundo ponto de reflexão deste texto: o encapsulamento da complexidade.

2.2. Encapsulamento da complexidade

Para que um sistema tenha viabilidade em sua utilização, um dos fatores principais é a curva de aprendizagem requerida aos seus usuários. Um exemplo bastante ilustrativo desta questão é a adesão à internet, que tinha em seus primeiros anos de acesso ao público, uma grande complexidade de uso, exigindo que um simples acesso a um determinado conteúdo, utilizasse comandos complexos, verdadeiras linhas de código, com informações, por exemplo, sobre o endereço e operação, sempre com alto grau de formalismo sintático. Esse modelo mantinha o grande público distante e apenas 'iniciados' na tecnologia se arriscaram a utilizá-la. Essa barreira foi quebrada com a proposta do modelo Web, que ocultava tais desdobramentos necessários às operações, por meio de interfaces gráficas que permitiam que uma simples ação do usuário, no recente e até então pouco utilizado mouse, fosse convertido 'internamente' nos complexos comandos executados pela máquina.

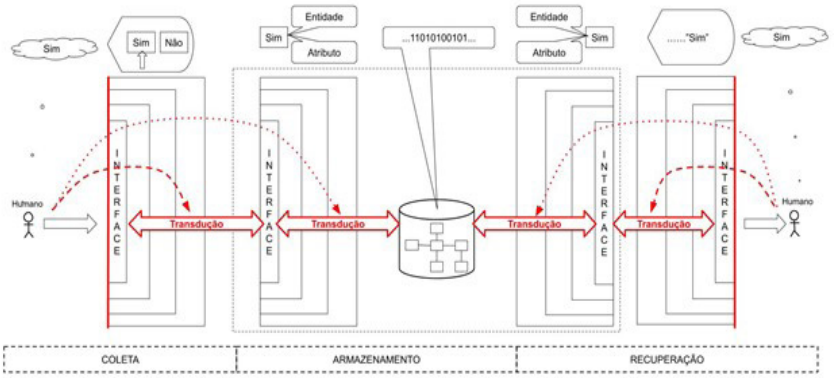
Este mesmo processo de ocultamento ocorre em todas as esferas de uso da tecnologia e, assim, esse encapsulamento da complexidade (Figura 3), permite que o utilizador possa se concentrar somente nos elementos estritamente necessários para sua interação com os sistemas. Se retomarmos ao exemplo apresentado na Figura 1 podemos inferir que o responsável pela digitação dos dados no sistema restringe seu foco ao conteúdo dos dados, sem se dar conta da forma fragmentada com que eles serão convertidos em entidades, e menos ainda, sobre como serão tratados fisicamente nos suportes digitais. Informações como a área do disco rígido onde será gravada a informação ou como a memória do dispositivo irá tratar estes conteúdos, ou mesmo como os diferentes programas irão interoperar, tais como o Sistema de Faturamento e o Sistema Gerenciador de Banco de Dados (Figura 3).

Este processo avançou tanto que hoje temos a viabilidade de interação direta do cliente com os sistemas de interface das empresas, o *e-commerce,* que aliando a facilidade de uso e ubiquidade da internet, atrelada à evolução dos sistemas de interface, excluem a participação daqueles que até então eram responsáveis pela venda e digitação (*input*) dos dados nos sistemas.

O encapsulamento da complexidade leva a uma percepção sobre os conteúdos, tratados pelos sistemas, totalmente baseada nas visualizações dos dados, sob forma de documentos ou relatórios, sempre voltados à busca pela aderência à necessidade informacional, ocultando, portanto, as estruturas de entidades utilizadas e, mais ainda, as formas de vinculação que permitem os relacionamentos entre estas enti-

dades de dados. Essa insciência do usuário sobre as transduções informacionais os distanciam de eventuais ações e definições necessárias para que a interoperabilidade seja viabilizada.

Figura 3 - Insciência do usuário sobre os processos de Transdução Informacional



Fonte: Elaborado pelo autor.

3. Reflexões sobre a análise e desdobramentos futuros

Como parte dos princípios FAIR, a ampliação da interoperabilidade, tão relevante para o pleno acesso a dados, é profundamente prejudicada pela insciência de atores envolvidos no processo das diversas transduções físicas e lógicas entre as fases de coleta, armazenamento e recuperação dos dados.

Reforça-se aqui a difícil percepção, por parte dos envolvidos, sobre a natureza estruturalmente fragmentada dos dados e da necessidade de ordená-los em conjuntos, que por sua vez, trazem suas próprias características para camadas de abstração que incorporam semântica a esses dados e permitem, finalmente, sua interpretação.

Mapeamentos lógicos, nas sucessivas camadas, incorporam dados sobre os dados (metadados), que precisam, entre outras finalidades, permitir a interpretação por humanos, e cada vez mais, por tratamentos maquínicos, o que leva à necessidade de compartilhamento de padrões físicos, lógicos e de fundo semântico complexos.

Dados coletados, em suas respectivas ambiências, recebem tratamento e são preparados para que sejam armazenados, prevendo, na maioria das vezes, seus usos dentro do contexto estabelecido no momento da coleta. No entanto, seu uso tende a ser cada vez mais disseminado e faz-se necessário que essas camadas se-

mânticas, agregadas aos dados, possam ser utilizadas por contextos não previstos ou até mesmo inexistentes no momento da coleta e armazenamento. Por outro lado, fatores como os relacionados com possíveis limitações para esses dados também requerem que essas aberturas, para usos não previstos, estejam explícitas, não só para os detentores dos recursos envolvidos no ciclo de vida dos dados, mas também aos usuários e eventuais referenciados por esses dados.

A Ciência da Informação pode, e deve, participar deste processo de identificação de fatores envolvidos no encapsulamento das transduções informacionais necessárias no processo de acesso a dados, e colaborar, não só na busca pela melhoria e ampliação do potencial de integração dos conteúdos informacionais contidos nos dados, como ainda, e principalmente, contribuir para a disseminação, junto à sociedade, do potencial uso de dados que, uma vez agregados e atendendo a requisitos de interoperabilidade, podem representar um valor muito maior que aqueles representados pelos conjuntos quando considerados individualmente.

4. Referências

DYCHÉ, Jill; LEVY, Evan. **Customer Data Integration**: reaching a single version of the truth. Hoboken, New Jersey: John Wiley & Sons, 2006.

FORCE11. **Guiding Principles for Findable, Accessible, Interoperable and Reusable Data Publishing** version b1.0. Disponível em: https://www.force11.org/fairprinciples Acesso em: 01 mai. 2020.

KELLEHER, John D,; TIERNEY, Brendan. **Data Science**. The MIT Press Essential Knowledge Series. Cambridge: The MIT Press, 2018.

REEVE, April. **Managing Data in Motion**: Data Integration best practice techniques and technologies. Waltham, EUA: Elsevier, 2013

SANT'ANA, R. C. G. **Tecnologia e gestão pública municipal**. São Paulo: Cultura Acadêmica, 2009. (Coleção PROPG Digital - UNESP). ISBN 9788579830105. Disponível em: http://hdl.handle.net/11449/109104

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. Informação & Informação, [S.l.], v. 21, n. 2, p. 116–142, dez. 2016. ISSN 1981-8920. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940>. Acesso em: 29 dez. 2016. doi:http://dx.doi.org/10.5433/1981-8920.2016v21n2p116.

SANT'ANA, Ricardo César Gonçalves. **Transdução Informacional**: impactos do controle sobre os dados. In: MARTÍNEZ-ÁVILA, D; SOUZA, E.A.; GONZALEZ, M. E. Q. (Orgs) Informação, conhecimento, ação autônoma e big data : continuidade ou revolução? Cultura Acadêmica - FiloCzar - São Paulo. 2019. p.117-128 ISBN 978-85-7249-054-2

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. **Dado e Granularidade na perspectiva da Informação e Tecnologia**: uma interpretação pela Ciência da Informação. Ciência da Informação, Brasília, v. 42, p. 199-209, 2015.

SHKEDI, Asher. Introduction to **Data Analysis in Qualitative Research**: practical and theoretical methodologies with use of a software tool. : SHKEDI, 2019.

WILKINSON Mark.D.; DUMONTIER Michel; MONS, Barend . et al. **The FAIR Guiding Principles for scientific data management and stewardship**. Sci. Data, 15; 3, 160018., 2016. Disponível em https://doi.org/10.1038/sdata.2016.18 Acesso em: 10 mai.2020.

Como citar com o DOI individual

SANT'ANA, Ricardo César Gonçalves. Interoperabilidade de dados e a transdução informacional encapsulada no acesso a dados *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 147 - 156. DOI: 10.22477/9786589167242.cap11

Desenvolvimento e aplicação de normas para interoperabilidade de repositórios de dados científicos: repositórios do IBICT e do CNPq

Lucas N. Paganine[1], Washington L. Ribeiro de Carvalho Segundo[2], João L. R. Moreira[3]

1. Introdução

Ciência Aberta (CA) é um termo guarda-chuva, que abrange diversos aspectos, como Ciência Cidadã e compartilhamento de cadernos de pesquisa. Ela surge com o desenvolvimento do Movimento Acesso Aberto (AA), tratando da abertura dos processos científicos para a população de uma forma geral, seguindo assim princípios de colaboração e transparência. No contexto da CA é ressaltado o crescimento da demanda do compartilhamento dos dados científicos, utilizando para este fim diversas ferramentas, entre elas destacam-se os repositórios de dados científicos.

Esses repositórios são ferramentas para tratamento, organização, disseminação e preservação de seus objetos digitais, neste caso os dados científicos. Contudo, devido às necessidades das diversificadas áreas do conhecimento e das diversas realidades institucionais que implementam os repositórios, surgem muitos padrões de descrição dos conjuntos de dados armazenados.

Destaca-se a importância de uma pesquisa acerca desses diferentes padrões e modelos existentes, visando o desenvolvimento de um esquema descritivo central que permita também o atendimento de necessidades específicas às diferentes áreas do conhecimento, via extensões deste padrão central, viabilizando assim a interoperabilidade entre repositórios de diferentes domínios temáticos.

1 Bacharel em Biblioteconomia com pós-graduação em Gestão Pública, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) - lucaspaganine@ibict.br

2 Doutor em Informática, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) - washingtonsegundo@ibict.br

3 PhD em Computer Science, Universidade de Twente - j.luizrebelomoreira@utwente.nl

Este estudo foi então desenvolvido no âmbito da OGP, uma iniciativa internacional iniciada em 2011 com o objetivo de incentivar a transparência como prática governamental, em especial, o acesso a informações públicas e a cooperação ativa com a sociedade. Objetivo este em grande consonância com a CA.

A OGP age por meio de Planos de Ação Nacionais com compromissos em práticas de Governo Aberto. Como resultado da execução destes planos são elaborados relatórios que expressam o andamento do atendimento das metas propostas.

Em 2018, o Brasil desenvolveu seu 4º Plano de Ação Nacional, com 11 compromissos, entre eles o Compromisso 3, que teve como objetivo "Estabelecer mecanismos de governança de dados de pesquisa para o desenvolvimento da Ciência Aberta no Brasil" (RNP, 2018). Este compromisso, conhecido como Compromisso pela CA, estava sob coordenação da Embrapa, mas com a participação de diversas instituições, em sua maioria governamentais.

O compromisso foi organizado em nove marcos, sendo o Marco 8 descrito como a "Proposição de padrões de interoperabilidade para repositórios de dados científicos" (RNP, 2018), coordenado pelo Ibict, mas com a colaboração da Rede Nacional de Ensino e Pesquisa (RNP), Universidade de Twente, Comissão Nacional de Energia Nuclear (CNEN) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Um dos resultados do Marco foi um Guia com a proposta de "Padrões de interoperabilidade para repositórios de dados de pesquisa" que são aplicáveis a qualquer repositório de dados de pesquisa que deseje promover a interoperabilidade e abertura dos dados científicos armazenados. São definidos no documento os critérios de interoperabilidade, orientando a construção ou aprimoramento de repositórios de dados científicos.

Tendo esse cenário em vista, vale explorar mais a fundo o produto do marco em questão, em especial ao considerar seu objetivo de "desenvolver e aplicar um conjunto mínimo de descrição para dados científicos, realizando-se apêndices para domínios específicos do conhecimento, com base nos padrões e diretrizes internacionais existentes" (PAGANINE; et al. , 2020).

Para tanto a metodologia aqui utilizada será iniciada com a descrição do desenvolvimento do resultado em questão bem como apresentação de seus resultados, e ao fim é fornecida uma descrição dos estágios e processos de aplicação deste documento nos repositórios de dados científicos do Ibict e do CNPq.

O documento em Paganine; et al. (2020) está dividido em 2 partes, uma com um conjunto de metadados gerais, e a outra com apêndices para áreas específicas do conhecimento. São utilizados como principais referências documentos de diretrizes internacionais bem estabelecidas para repositórios de dados científicos. Foram elas: as Diretrizes OpenAIRE para Repositórios de Dados, concomitantemente às

Diretrizes OpenAIRE para Repositórios de publicações científicas; e o conjunto de metadados descrito pelo *framework Fair Data Point* (FDP). As diretrizes OpenAIRE são extensivamente adotadas internacionalmente, porém, pesquisas tratando da interoperabilidade semântica indicam a importância de extensões para atender os princípios FAIR (WILKINSON; et al, 2016). Esta extensão citada também por Santos; et al. (2016) é explorada no *framework* FDP.

Para a classificação acerca das extensões à áreas do conhecimento específicas, por questões organizacionais internas ao Brasil, foi utilizada inicialmente para comparação as tabelas de áreas do conhecimento CNPq (CNPQ, s.d.), a de áreas do manual de pesquisa Frascati (OECD, 2015) e a da divisão de áreas utilizada pelo Data Curation Centre (DCC) e RDA, a *Deutsche Forschungsgemeinschaf* (DFG, 2020). A tabela comparativa completa é encontrada no documento original resultado do marco.

Sobre o levantamento de padrões para as extensões para áreas do conhecimento específicas, este iniciou-se com o Diretório de metadados[4], ferramenta mantida pela RDA, seguida de subsequente análise de padrões de metadados utilizados em repositórios temáticos (abrangem apenas uma dada área do conhecimento) e institucionais e/ou multitemáticos de dados encontrados no Registry of Research Data Repositories (re3data).

Ao final realizou-se a conferência dos padrões encontrados nos repositórios levantados, em busca de quais apêndices possam complementar a OpenAIRE somada do padrão FDP para as 4 áreas inicialmente selecionadas: Biologia (devido ao seu comportamento nas árvores do conhecimento levantadas), Agricultura (devido à importância da área nacionalmente), Saúde (pela importante e pioneira atuação no movimento de Acesso Aberto e outros aspectos relacionados à pesquisa) e Ciências Sociais (devido ao objeto de estudo geral do instituto onde a pesquisa foi desenvolvida, o Ibict). Vale destacar que durante o desenrolar da proposta especialistas das áreas foram consultados, em especial da Fiocruz (Saúde) e da Embrapa (Agricultura).

2. Desenvolvimento

As diretrizes presentes em Paganine; et al. (2020) tomam também como referência as antigas Diretrizes DRIVER que foram publicadas em 2007 pelo projeto *Digital Repository Infrastructure Vision for European Research (*DRIVER*)* e as *Guidelines for content providers: Exposing textual resources with* OAI-*PMH*, contendo recomendações iniciais que para interoperabilidade. Estas recomendações foram comple-

4 rd-alliance.github.io/metadata-directory/

mentadas pelas *OpenAIRE Guidelines for Literature Repositories*, que no momento da elaboração do presente texto se encontravam na versão 4 *(*OPENAIRE, 2018*)*.

As diretrizes da OpenAIRE estão organizadas em três seções: A primeira é introdutória; Na segunda, se descreve o uso do OAI-*PMH (Open Archives Initiative Protocol for Metadata Harvesting)*, com orientações sobre este e; finalmente uma visão geral de um perfil para aplicação. Esta diretriz é composta por 4 padrões de metadados: *Dublin Core;* e sua versão qualificada; *Datacite* e; Oaire (padrão elaborado pela própria *OpenAIRE*). Também são especificados alguns vocabulários controlados para uso, como por exemplo os da *Confederation of Open Access Repositories (*COAR*)*.

Sobre o protocolo OAI-PMH, este é uma ferramenta para a exposição de metadados através das linguagens *Hypertext Transport Protocol (*HTTP*)* e *Extensible Markup Language (*XML*)* que permite a comunicação e interoperabilidade entre bases. Vale destacar que independentemente de seu grande uso na interoperabilidade entre sistemas, e de sua recomendação nas nas diretrizes da OpenAIRE, percebe-se já uma certa limitação desta ferramenta, em seu conjunto de metadados (15 elementos do *Dublin Core*), que vem sendo utilizada desde o início dos anos 2000 (GARCIA; SUNYE, 2003). Outras iniciativas estão cada vez mais proeminentes, em especial sobre Interoperabilidade Semântica, como o W3C: PROV-O (modelo de dados genérico da *World Wide Web Consortium*) e o Data Catalog Vocabulary (DCAT) 2.0 (um vocabulário em Resource Description Framework- RDF), que adiciona classes para descrição de serviços de dados voltando-os para os princípios FAIR.

De volta às diretrizes da OpenAIRE V4, nelas são estabelecidos 4 níveis de obrigação de preenchimento de campos: Mandatório (M), quando o preenchimento é obrigatório (aplicado a 6 campos); Mandatório se aplicável (MA), se o preenchimento for obrigatório apenas no caso do campo ser informação parte do registro (exemplo, o nome do organismo financiador é obrigatório, caso o conjunto de dados tenha sido fruto de um financiamento) que é aplicado à 8 campos; Recomendado (R), que possui relevância e importância mas não é essencial (aplicado a 15 campos) e; Opcional (O), que apenas acrescentaria valor à descrição, mesmo sendo desnecessário (aplicado a 3 campos). O perfil de aplicação da diretriz é, em suma, representado pela tabela que relaciona os trinta e dois campos da diretriz, com as orientações de preenchimento e os vocabulários controlados selecionados para campos específicos.

A segunda ferramenta em análise é o FDP, um aplicativo web de código aberto e independente, desenvolvido como implementação de referência das especificações

do próprio FDP[5]. Estas especificações guiam softwares para repositórios, tratando da gestão de metadados, em particular sobre tecnologias semânticas como o RDF, sendo assim uma ferramenta complementar a um software de repositórios de dados. Um repositório baseado no FDP trabalha questões de interoperabilidade, possibilitando a encontrabilidade, acessibilidade, interoperabilidade e reutilização (os princípios FAIR).

A implementação de referência do FDP utiliza uma API REST com diversas funções: criação, armazenamento e veiculação de metadados permitindo assim a exposição, fornecimento e disponibilização destes metadados de forma adequada aos princípios FAIR. Ele também permite o descobrimento de metadados de conjuntos disponibilizados e o acesso a estes, quando estão em uma licença de uso aberta. Qualquer repositório de dados pode adotar os metadados do FDP, funcionando assim também como uma instância FDP.

Uma das principais especificações do FDP é a especificação de níveis de metadados[6], que orienta a aplicação de um perfil em RDF, o qual reutiliza modelos semânticos padronizados. Portanto, a especificação de níveis de metadados do FDP introduz uma organização em quatro níveis de metadados: primeiro, o próprio repositório de metadados; segundo, o catálogo; terceiro, o conjunto de dados e; quarto, a distribuição de dados (que são os arquivos pertencentes ao conjunto). Um repositório de metadados pode possuir um ou mais catálogos, cada catálogo pode possuir um ou mais conjuntos de dados, e cada conjunto de dados contém uma ou mais distribuições.

O padrão de metadados FDP tem por sua vez base no Esquema re3data[7] e no vocabulário DCAT[8]. Como observado anteriormente, o padrão de metadados é organizado em quatro níveis e cada propriedade tem duas possibilidades de preenchimento: Obrigatório, aplicado a dez campos, e Opcional, aplicado a doze campos. Os níveis descrevem, cada um, um tipo de objeto digital complexo que possivelmente é descrito, são elas: o nível de repositório de metadados, contendo informações sobre o repositório de dados e do próprio FDP; o nível de metadados do catálogo, contendo informações acerca da coleção, onde cada catálogo representa uma categoria (geralmente definida por domínio; o nível de metadados do conjunto de dados, contendo informações sobre o conjunto de dados em si; e o nível de metadados de distribuição, contendo informações sobre as possíveis se-

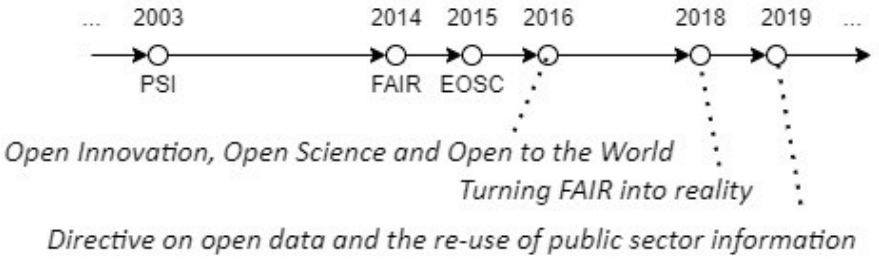
5 github.com/FAIRDataTeam/FAIRDataPoint-Spec
6 github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/spec.md
7 https://www.re3data.org/schema
8 https://www.w3.org/TR/vocab-dcat-2/

rializações do conjunto de dados, i.e., os arquivos individuais que compõem os conjuntos de dados.

Por exemplo, o repositório de dados B2Share (https://b2share.eudat.eu/) aborda catálogos através de comunidades. A Kinder Corona Studies (KiCoS) é uma das comunidades (catálogos) do B2Share e contém uma série de conjuntos de dados (na ferramenta representado como *records*); e um conjunto de dados pode conter uma série de distribuições (*files*). A implementação da especificação de metadados do FDP como um "*proxy* (*wrapper*) semântico" pode agregar as funcionalidades supracitadas ao software de repositórios de dados (Moreira et al, 2019).

Destacam-se ainda no cenário europeu ações e programas da Comissão Europeia (CE) que tratam do compartilhamento e abertura de dados científicos, tais como: a Nuvem Europeia de Ciência Aberta (EOSC), que data de 2015; a evolução dos princípios FAIR, que têm como ponto de partida o ano de 2014; Em 2016, publica-se o documento Inovação aberta, ciência aberta e aberta ao mundo: uma visão para a Europa; Em 2018, acontece a publicação do relatório Transformando o FAIR em realidade: Relatório Final e Plano de Ação sobre Dados do FAIR, e percebe uma crescente participação da CE na RDA; Já em 2019, ocorre a transformação das diretrizes Informações do Setor Público (PSI), que haviam sido editadas em 2003, nas Diretrizes para dados abertos e reuso de informações do setor público (vide Figura 1, abaixo).

Figura 1- Linha do tempo outras iniciativas CE



Fonte: Elaborado pelo(a) autor(a).

A elaboração das diretrizes gerais para repositórios de dados científicos inicia-se com uma análise comparativa entre os conjuntos OpenAIRE e FDP em busca de diferenças e similaridades entre os requerimentos. A comparação completa pode ser encontrada no documento original resultado do marco.

Percebeu-se a possibilidade de equivalência entre grande parte dos campos comparados, destaca-se apenas as diferenças nas definições dos níveis de obrigato-

riedade dentro dos padrões acerca de campos equivalentes. Partindo do esclarecimento destas diferenças de obrigatoriedade, iniciou-se a busca pela definição dos metadados mínimos obrigatórios.

Foram encontradas dificuldades na definição de equivalências, em especial em campos OpenAIRE relacionados aos subtipos *Type* (como por exemplo *dateType*), além da adoção de diferentes vocabulários controlados. Em particular, a OpenAIRE recomenda um vocabulário controlado para *resourceType*, chamado *Controlled Vocabulary for Resource Type Genres (Version* 2.0*)*, que é uma taxonomia de classificação de tipologias de gênero de recursos. Decidiu-se pela não adoção desta taxonomia pois identificou-se que ela apresenta uma série de problemas semânticos na sua relação hierárquica, uma vez que não é possível identificar qual tipo de relação é usada, por exemplo se é uma relação de especialização, como o *rdf:Type* (ou *"is a"*), ou se é outro tipo de relação. Um outro exemplo é o caso em que a taxonomia descreve que uma entrevista (*interview*) é um conjunto de dados (*dataset*), o que não parece ter sentido uma vez que uma entrevista é uma ação intencional (um evento, ou *perdurant*), enquanto um conjunto de dados é uma substância (um *endurant*) que pode ter diferentes princípios de identidade. Esta taxonomia também apresenta problemas de fundamentação em relação aos princípios de identidade, rigidez e disjunção lógica das categorias. Por exemplo, a taxonomia apresenta no mesmo nível os elementos *learning object* e *text*, os quais podem ser (ou não) disjuntos, e não compartilham o mesmo princípio de identidade.

É importante salientar que ao adotar uma taxonomia deste tipo pode-se provocar uma série de dificuldades na interoperabilidade, uma vez que as máquinas necessitam de descrição precisa das tipologias de recursos disponíveis nos repositórios de dados. Encontrou-se também incompatibilidade no campo *dcat:distribution* do FDP, esse campo pede uma descrição de informações acerca do arquivo individual que compõe o conjunto de dados (*dcat:Dataset*).

O padrão central mínimo obrigatório desenvolvido abrange 13 campos com exemplos de preenchimento e aplicação que podem ser encontrados no documento original resultante do marco em Paganine et al. (2020).

Partindo-se da definição do núcleo, é abordado o desenho dos metadados temáticos. Uma dificuldade encontrada foi a questão da multidisciplinaridade frequente mesmo em repositórios monotemáticos. Para tanto utilizou-se o quadro de comparação das áreas do conhecimento como norteador na escolha e organização dessas áreas.

Iniciou-se pela procura por esquemas de metadados temáticos. Como ferramenta inicial foi utilizada uma página publicada pela DCC em 2020, com diferentes esquemas metadados, divididos por campos do conhecimento. Esta lista foi então

analisada e deduplicada para os padrões das áreas escolhidas (Ciências sociais, Biologia e Agronomia). Porém o resultado obtido não apresentou especificidade e abrangência satisfatórias. Foi então utilizada uma lista de padrões mantida por um grupo de interesse em metadados da RDA[9] (RDA, 2020). Por fim, realizou-se um refinamento do resultado obtido, com checagem dos metadados obtidos, com a lista exibida no registro re3data[10], aferindo-se quais padrões dos selecionados são efetivamente utilizados em repositórios de dados temáticos afetos às áreas selecionadas.

Seguindo o nível de complexidade do mais simples ao mais difícil de ser tratado, seguiu-se pela análise das tabelas da área de Biologia. Foram analisados três esquemas de metadados: o MIBBI, Darwin Core e ABCD. O MIBBI conta com 40 categorias ou como são chamados, módulos, com 23 campos principais, mas apenas 17 destes já estão completamente desenvolvidos, sendo o restante ainda em elaboração, no momento em que a tabela foi coletada. O Darwin Core tem 12 pacotes de descrição com carga semântica ou categorias, enquanto o ABCD conta com 38 categorias de metadados expansíveis. Durante a navegação no re3data percebe-se, excluindo esquemas genéricos e de outras diferentes ou gerais áreas do conhecimentos relacionadas, que dos 42 resultados, 2 utilizam MIBBI e apenas uma o Darwin Core.

Após a análise comparativa, foram selecionados os metadados de aparente importância na área, de acordo com sua frequência nos padrões, eliminando redundâncias e generalizando os termos similares. O resultado também é encontrado no documento original , contendo 8 campos, utilizando-se o Darwin Core. A qual foi escolhida pela facilidade de tradução e comparação com o tradicional Dublin Core:

Ao se dar foco às Ciências Sociais, que é uma área de maior abrangência nos padrões escolhidos, também se constata uma abrangência de padrões. São eles: METS, MODS, MARC, CERIF e Dublin Core (DC). O METS conta com 7 classes, já o MODS, 20, o DC, 15, o MARC, 9 e o CERIF, com 22. Essas classificações e respectivos campos ou elementos dos padrões foram então comparados seguindo o mesmo processo aplicado nos padrões de Biologia. O resultado selecionado foi então adaptado na linguagem Datacite, pelo seu desenvolvimento já voltado para dados científicos e a extensa aplicação e compatibilidade.

Por fim, ao se tratar a área de Agricultura, segue-se o mesmo processo aplicado em Biologia. Foram eleitos os esquemas AGRIS e AgMES. O AgMES tem 21 campos que se baseiam no DC e abrangem padronizações semânticas na agricultura sobre descrição, descobrimento de recursos, interoperabilidade e intercâmbio de

9 rd-alliance.github.io/metadata-directory/

10 https://www.re3data.org/

dados em diversos recursos informacionais. Já o AGRIS conta com 16 campos e é voltado para o sistema internacional de informação sobre diretrizes de ciências e tecnologias agrícolas, sobre boas práticas para a informação. No registro re3data não se identificou o uso declarado de nenhum dos 2 esquemas em repositórios em repositórios temáticos da área de agricultura. Para descrição dos campos aqui, foi escolhido o esquema AgMES, pela proximidade com o DC.

Destacam-se também os resultados negativos obtidos sobre Saúde, a área apresentou uma alta complexidade não prevista de padrões de descrição de conjuntos. Deparou-se com uma ausência de documentação sobre estes, não sendo assim identificados esquemas de metadados específicos à área. Na tentativa de capturar alguns dos campos utilizados com frequência na área, foram realizadas consultas aos repositórios de dados temáticos e a alguns especialistas da Fiocruz, porém os resultados obtidos ainda não foram satisfatórios e suficientes para a elaboração de uma proposta que englobasse toda área de Saúde.

O padrão de metadados geral multitemático desenvolvido por este trabalho já está em aplicação no desenvolvimento dos repositórios de dados científicos do Ibict (denominado *Aleia*) e do CNPq (denominado *LattesData*).

As criações do Aleia e do LattesData também foram motivadas pelo compromisso 3, do 4º Plano de Ação Nacional da OGP Brasil, através de um acordo de cooperação técnica (ACT) firmado entre CNPq e Ibict, em dezembro de 2019. O Aleia tem por objetivo prover uma ferramenta com as funcionalidades de registrar, reunir, organizar, disseminar, compartilhar e preservar os dados científicos de pesquisas executadas por colaboradores do Ibict e conjuntos de dados científicos externos ao órgão, mas de comunidades científicas específicas. Já o LattesData tem por objetivo principal servir como ferramenta oficial que possibilite seus pesquisadores financiados a realizarem depósitos de conjuntos de dados que surgiram como resultado dos projetos desenvolvidos com recursos do CNPq, fazendo parte dos procedimentos de prestação de contas, bem como às instituições não clientes e parceiras do CNPq que firmem acordos para uso colaborativo do espaço criado.

Os 2 repositórios serão multitemáticos e pretendem abranger conjuntos de dados de pesquisadores de diversas instituições, de diversas áreas e realidades. Com isto em mente, decidiu-se iniciar apenas pela aplicação do padrão de descrição central mínimo com apenas algumas pequenas alterações e adições para melhor se adequar a sua instituição mantenedora.

Os metadados adicionais do Aleia são apresentados a seguir.

Quadro 1- Metadados adicionais Aleia

Campo	Descrição
Currículo Lattes do autor	Endereço para acessar o Currículo Lattes (este campo foi apenas adaptado a partir do campo "Identificador" do padrão original)
Instituição de origem do autor	Nome completo da instituição à qual o pesquisador esteja vinculado
Descrição	Descrição geral do conjunto de dados e seu conteúdo
Título alternativo	Título do conjunto de dados em outro idioma alternativo
Contato	Endereço de correio eletrônico (e-mail) do responsável pelo conjunto de dados
Idioma do conjunto de dados	Idioma no qual o conjunto de dados foi desenvolvido
Software de leitura e manipulação de dados	Programa utilizado para acessar e manipular os arquivos do conjunto dados

Fonte: Elaborado pelos autores

Os metadados adicionais do LattesData adicionados são apresentados no quadro a seguir.

Quadro 2: Metadados adicionais LattesData

Campo	Descrição
Contato	Endereço de e-mail preferencialmente institucional do responsável para contato sobre os dados
Identificador interno de autor (IDLattes)	Definição: Número identificador do currículo Lattes (este campo foi apenas adaptado a partir do campo "Identificador" do padrão original)
Identificador externo de autor	Número identificador do ORCID
Instituição do autor	Nome da instituição a qual o autor está filiado
Identificador alternativo do conjunto de dados	Outro identificador persistente do conjunto de dados obtido de outra forma que não o principal utilizado no repositório LattesData
Notas	Texto livre podendo ser utilizado para listagem\descrição dos arquivos (relacionar cada arquivo, seu tipo, descrição e caso necessite de software específico também informar), comentário ou orientações de acesso e outros detalhes
Resumo do projeto	Texto explicativo que descreve o projeto e conjunto de dados de forma geral, podendo abranger conclusões, metodologia, coleta, etc
Valor recebido	Valor em reais recebido pelo projeto/chamado do financiador
Vigência do projeto	Data de início e final do projeto
Materiais ou outros produtos relacionados	Qualquer produto ou material relacionado ao conjunto de dados que não publicação científica formal padrão

Fonte: Elaborado pelos autores

Definidos os campos adicionais em ambos os repositórios, encontraram-se grandes dificuldades de alteração destes e do formulário no software eleito (Harvard Dataverse). Como tentativa de solução, está em desenvolvimento um formulário de comunicação com a API REST do Dataverse, para recuperação e preenchimento de metadados de forma externa ao software. Outra dificuldade encontrada

é a alteração dos prefixos dos campos que pelo software vem com o padrão da Data Documentation Initiative (DDI) ao invés do DCAT e esquema Datacite recomendados. Futuramente planeja-se integrar o preenchimento dos metadados com a geração de um Plano de Gestão de Dados (PGD), para que este tenha também um formato legível por máquinas.

Ao analisar os objetivos e resultados até então alcançados, percebe-se que o núcleo geral de metadados mínimos apresenta as informações suficientes para interoperabilidade das diretrizes eleitas. Porém é interessante que se realce a importância da adição de outras informações de relevância institucional para contribuir com uma descrição de maior qualidade dos conjuntos depositados, e sua associação aos projetos financiados, no caso do CNPq.

Por fim também é evidenciado a magnitude da complexidade do tratamento e descrição de diferentes áreas do conhecimento, em repositórios multitemáticos e em especial ao se considerar a possibilidade de abranger também instituições com realidades e contextos muito diversificados.

3. Conclusão

A principal dificuldade enfrentada na execução deste trabalho foi o fato dos repositórios não adotarem padrões de metadados bem difundidos, e a constatação de que os padrões adotados em diferentes áreas não apresentam compatibilidade entre si. O conteúdo apresentado como resultado do trabalho executado na OGP se apresenta adequado ao tratar áreas que possuem esforços em direção à representação de produtos científicos, mas a descrição dos conjuntos de dados, que potencialmente abrange qualquer área do conhecimento, mostra-se um trabalho que se encontra em constante mudança e atualização. Destaca-se então que este não é um estudo que termina em sua aplicação, ao contrário, necessita de constante desenvolvimento em busca de extensões e adequações para outras áreas do conhecimento. Pretende-se também elaborar um *corpus* composto por um conjunto de metadados que descrevem dados de pesquisa para cada área específica do conhecimento. Essas informações serão coletadas nos repositórios de dados de temas únicos cadastrados no *re3data* que possibilitam a comunicação. A coleta será automatizada e abrangerá: título, palavras-chave, resumo e assunto. O *Corpus* será então organizado e sua visualização gerada com o programa VosViewer para identificação de assuntos-chave e áreas específicas altamente populadas por conjuntos de dados.

Destaca-se que o trabalho apresentado foi elaborado em colaboração com o IBICT, o CNPq e a Universidade de Twente. Para futuros desenvolvimentos pretende-se continuar estudos em outras áreas do conhecimento, em especial as ciências exatas e saúde, assim como na análise ontológica de tipologias de gênero de recur-

sos para endereçar os problemas identificados na classificação recomendada pelo OpenAIRE. As atividades futuras também incluem adoção abrangente dos princípios FAIR e, particularmente, a evolução da interoperabilidade semântica entre repositórios de dados por meio da aplicação de ontologias bem fundamentadas nos repositórios em desenvolvimento.

4. Referências

CNPQ. **Tabela de Áreas do Conhecimento**. Brasília: CNPQ, 2020. Disponível em: http://lattes.cnpq.br/web/dgp/arvore-do-conhecimento. Acesso em: 28 mai. de 2020.

DCC. **Disciplinary Metadata**. [*S.l.*], 2020 .Disponível em: dcc.ac.uk/resources/metadata-standards. Acesso em: 28 mai. 2020.

DFG. DFG **Classification of Subject Area, Review Board, Research Area and Scientific Discipline**. Bonn, 2020. Disponível em: dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp. Acesso em: 28 mai. de 2020.

FDP. FAIR **Data Point Specification**. [*S.l.*], 2016. Disponível em: github.com/FAIRDataTeam/FAIRDataPoint/wiki/FAIR-Data-Point-Specification. Acesso em: 28 mai. de 2020.

GARCIA, Patrícia de Andrade Bueno; SUNYE, Marcos Sfair. O protocolo OAI-PMH para interoperabilidade em Bibliotecas Digitais. In: **Congresso de Tecnologias para Gestão de Dados e Metadados do Cone Sul**. 2003.

MOREIRA, J. L. R.; BONINO, L.; FERREIRA PIRES, L.; VAN SINDEREN, M.; HENNING, P. Repositórios para dados localizáveis, acessíveis, interoperáveis e reutilizáveis (FAIR): adaptando um repositório de dados para se comportar como um FAIR Data Point. **Liinc Em Revista**, 15 (2). 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4817. Acesso em: 28 mai. de 2020.

OPENAIRE. DRAFT: OpenAIRE Guidelines for Literature Repository Managers v4. [*S.l.*], 2018. Disponível em: openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/. Acesso em: 28 mai. de 2020.

OECD. **Frascati Manual** 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities. Paris: OECD Publishing, 2015. Disponível em: oecd.org/publications/frascati-manual-2015-9789264239012-en.htm. Acesso em: 28 mai. de 2020.

PAGANINE, Lucas. *et al.* **Padrões de interoperabilidade para repositórios de dados de pesquisa** Brasília: IBICT, 2020. 44 p. Disponível em: livroaberto.IBICT.br/bitstream/123456789/1085/2/Padr%c3%b5es%20de%20

interoperabilidade%20para%20reposit%c3%b3rios%20de%20dados%20de%20pesquisa%20OGP%20.pdf. Acesso em: 28 mai. de 2020.

RDA. **Metadata Directory**. [*S.l.*], 2020. Disponível em: rd-alliance.github.io/metadata-directory/. Acesso em: 28 mai. de 2020.

RE3DATA. **Registry of research data repositories**. [*S.l.*], 2020.Disponível em: service.re3data.org/about. Acesso em: 28 mai. de 2020.

RNP. Wiki Ciência Aberta na OGP Brasil. 2018. Disponível em: wiki.rnp.br/pages/viewpage.action?pageId=107315238. Acesso em: 30 abr. 2020.

SANTOS, L. FAIR Data Points Supporting Big Data Interoperability. *In*: MERTINS, K. *et al.* (org.). **Enterprise Interoperability in the Digitized and Networked Factory of the Future**. London: ISTE Press, 2016. p. 28 mai. de 2020.

WILKINSON. Mark D. *et. al*. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, [*S.l.*], v.3, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 28 mai. de 2020.

Como citar com o DOI individual

PAGANINE, Lucas N; CARVALHO SEGUNDO, Washington L. Ribeiro de; MOREIRA, João L. R. Desenvolvimento e aplicação de normas para interoperabilidade de repositórios de dados científicos: repositórios do IBICT e do CNPq. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 157 - 170. DOI: 10.22477/9786589167242.cap12

Investigando os princípios FAIR em repositórios de dados científicos do National Institutes of Health (NIH)

Marcello Peixoto Bax[1]

1. Introdução

A coleta e a análise de dados são essenciais para todas as ciências, mas são especialmente importantes quando se trata das ciências biomédicas ou da saúde. À medida em que o mundo avança no caminho em direção à medicina personalizada e à maior rapidez e agilidade na produção de medicamentos, a compreensão dos dados coletados de ensaios clínicos e outros estudos é crucial para acelerar o avanço da pesquisa científica. Infelizmente, devido à multiplicação de dados e formatos, a coleta, organização e disseminação de dados estão se tornando cada vez mais difíceis de realizar com eficiência. Além disso, compreender fenômenos e comprovar uma hipótese nesse campo requer grandes quantidades de dados, e poucos pesquisadores possuem recursos e meios para coletar tal quantidade de informações. Ensaios clínicos têm alto custo, requerem recursos não triviais e podem levar anos para serem realizados, dependendo do estudo. Obviamente, como tal, este tipo de coleta e aquisição de dados não está prontamente disponível para a grande maioria dos pesquisadores. É por isso que o campo da saúde, e outros domínios do conhecimento estão se movendo em direção ao compartilhamento generalizado desses dados por meio de *data centers* públicos ou privados disponíveis aos pesquisadores.

Infelizmente, como resultado de um gerenciamento de dados inadequado, são ainda muito escassas as iniciativas de compartilhamento de dados bem-sucedidas. O gerenciamento de dados é "o principal canal para a descoberta e inovação do conhecimento", promovendo o compartilhamento e a reutilização de dados em comunidades científicas (wilkinson *et al.*, 2016). Passou a ser importante definir um conjunto de princípios comuns que definem o que deve ser um "bom" gerenciamento de dados. Esses princípios, que destacam a capacidade de localização, acessibilidade, interoperabilidade e de reutilização dos conjuntos de dados (*data-*

1 Escola de Ciência da Informação ECI – UFMG – Pós-doutor no TWC - RPI NY, bax@ufmg.br

sets), são conhecidos como Princípios Orientadores FAIR. Tais princípios são cada vez mais considerados como uma referência para os *data centers* e estão sendo usados para avaliar e destacar o sucesso de certas iniciativas. Inúmeras publicações discutem a adesão aos princípios FAIR como uma forma de ilustrar o compromisso em facilitar o compartilhamento de dados em suas respectivas comunidades. Citam-se, por exemplo, o *Immune Epitope Database* (VITA *et al.*, 2018), a *DisGeNET Platform* (PIÑERO *et al.,* 2017), o *BioSharing Portal* (MCQUILTON *et al.,* 2016) e o *Omics Discovery Index* (PEREZ-RIVEROL *et al.,* 2017).

Existem vários problemas comuns que impedem os dados de serem considerados FAIR. Primeiro, poucos *datasets* podem ser harmonizados entre si; vários estão intimamente relacionados, mas os dados não são formatados da mesma maneira, portanto, não são facilmente harmonizáveis e não podem ser integrados para análise. A harmonização dos dados requer o uso de categorias e unidades de medidas comuns ou pelo menos comensuráveis. Em segundo lugar, no caso de pesquisas envolvendo diferentes equipes, o coordenador da pesquisa (investigador principal) geralmente sabe bem mais detalhes sobre a natureza estrutural dos dados que coletou (suas propriedades e relações), do que é capaz de transmitir de forma coesa como informações suplementares que acompanhariam os próprios dados (metadados). Finalmente, se a quantidade de dados for suficientemente volumosa, métodos automatizados podem ser a única maneira viável de gerar análises abrangentes e profundas sobre eles. No entanto, se os significados dos dados não forem formalizados explicitamente de maneira a se tornarem "legíveis por máquina", não haverá método automatizado que possa dar suporte a essa análise. É aqui que entra o conceito de "*lifting*" semântico de dados ou ainda de "ingestão" semântica de dados em repositórios.

2. *LIFTING* Semântico de dados

O *lifting (*elevação) semântico de dados é um processo pelo qual os dados são convertidos de sua representação original tabular, em arquivos CSVs e/ou tabelas relacionais, para uma representação formal ontológica que representa o "conhecimento" na estrutura de um grafo de conhecimento (PAN *et al.,* 2017). Nesta operação, os dados não são apenas convertidos para outro formato, mas "elevados" ao nível de "conhecimento" já que passam a ser representados por modelos ontológicos fundados em lógicas de descrição que explicitam a semântica formal dos mesmos. O processo transforma os dados, originalmente sem significado explícito, em dados potencialmente interoperáveis na web semântica (*linked data*) e tratáveis por computador. O *lifting* de dados é, portanto, importante porque ajuda a combater todos os problemas mencionados acima que são alvo de tratamento dos princípios

FAIR. Os dados são coletados e reestruturados em formato acessível, orientado por metadados e legível por máquina. Dados nesse formato podem ser mais amplamente divulgados pela web para posterior extração de informação e conhecimento, preservando seu significado original.

Diversos *data centers* estão trabalhando para aumentar o *FAIRness* de seus repositórios de dados. Em alguns casos, isso é feito com o desenvolvimento ou integração de plataformas de software que incorporam um processo de enriquecimento semântico dos modelos de dados. Algo que pode ser alcançado representando o modelo, ou parte dele, com um formalismo orientado por ontologias (*lifting* semântico) e mapeando-os para uma ontologia de referência. Como este é um fenômeno relativamente recente, não há um método consensual de realizar o processo de *lifting* e ingestão. Portanto, é importante entender o que diferentes organizações estão fazendo em sua tentativa de melhorar o estado do *lifting* semântico de dados como uma forma de aprender com cada um desses esforços.

3. Data centers financiados pelo *National Institutes of Health*

O *National Institutes of Health* (NIH) financia centenas de *data centers* diferentes nas áreas da saúde. Alguns desses contêm *data sets* únicos para um domínio específico, enquanto outros hospedam vários conjuntos de dados diferentes em vários domínios e agências do NIH. Embora o Instituto incentive os *data centers* a utilizar repositórios específicos de domínio sempre que possível, esses repositórios não estão disponíveis para todos os *data sets*. Quando os pesquisadores não conseguem localizar um *data center* que mantém repositório para sua disciplina ou para o tipo de dados que geram, um repositório generalista pode ser um lugar útil para compartilhar dados. Os repositórios generalistas aceitam dados independentemente do tipo, formato, conteúdo ou foco disciplinar. O NIH não recomenda um repositório generalista específico, mas mantém uma lista não exaustiva, fornecida como guia para localizar repositórios. A lista contém os seguintes repositórios generalistas mais conhecidos: *Dataverse, Dríade, Figshare, Mendeley Data, Open Science Framework, Vivli e Zenodo*. Uma relação abrangente de *data centers* de compartilhamento de dados financiados pelo NIH foi criada pela US *National Library of Medicine*, onde o *Trans*-NIH *Biomedical informatics coordinating committee* (BMIC)[2] mantém outra lista com atualmente 66 *data centers* que mantém repositórios de domínio específicos e abertos, financiados pelo NIH. Outros 31 específicos de domínio apoiados incluem aqueles que têm limitações no envio e/ou acesso aos dados (dados sensíveis).

2 https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Estudar todos os 97 repositórios desses *data centers* seria impraticável. Compreender os recursos técnicos de um repositório de dados não é trivial e geralmente requer acesso a pelo menos alguns dos dados disponíveis. Assim, para começar, **pesquisou-se cerca de** 10 **repositórios** cujas descrições se destacaram pelo seu nível de detalhamento. Esses repositórios foram então inspecionados quanto às capacidades técnicas que os diferenciassem de outros. A pesquisa revelou que alguns desses repositórios estão, na verdade, hospedados em plataformas de software desenvolvidas por terceiros que contém dados de vários estudos e instituições diferentes. Isso gera uma dinâmica interessante em que alguns *data centers* criam repositórios e hospedam seus próprios dados, enquanto outros simplesmente hospedam os dados para instituições interessadas. Por fim, três *data centers* foram selecionados para uma inspeção mais detalhada de seus repositórios: **ImmPort**, **Synapse** e NDA (*National Data Archive*) do *National Institute of Mental Health (*NIMH*)*. Esses três centros foram selecionados pela possibilidade de acesso aos dados. Cada um desses centros contém pelo menos alguns repositórios que permitem o acesso público aos dados resumidos, no mínimo. Ter acesso aos dados permitiu uma maior compreensão de como estes são armazenados e como podem ser buscados. Outra razão pela qual eles foram selecionados foi a sua usabilidade. Essas plataformas tinham mecanismos relativamente simples de busca de dados, que ilustravam claramente o potencial das buscas. Deve-se notar, entretanto, que muitos outros *data centers* sofisticados existem em vários outros países, inclusive no Brasil, e que esta análise não pretendeu excluir as suas relevantes contribuições. Contudo, as restrições de tempo e recursos desta pesquisa exigiram que algumas plataformas mais facilmente acessíveis fossem consideradas.

4. Critérios de avaliação

A análise desses *data centers* destacou quatro critérios de avaliação: 1) como os dados podem ser encontrados e buscados/consultados? 2) são dados de domínio único ou de múltiplos domínios cruzados? 3) o esquema de representação dos dados é livre (*schema-free*), ou fixo/relacional? 4) o repositório realiza o *lifting* dos dados ao fazer a sua ingestão em algum banco de dados? Esses critérios foram selecionados porque cada um deles serve como um indicador do grau de aderência dos dados aos princípios FAIR. O primeiro critério acima satisfaz certas características de localização e acessibilidade (*Findable* e *Accessible*), pois os recursos de filtragem/consulta exigem que os dados sejam "descritos com metadados ricos", "registrados ou indexados de forma pesquisável" e "recuperáveis por seu identificador usando um padrão protocolo de comunicações" (WILKINSON *et al.*, 2016). O domínio ou área de conhecimento de um repositório, segundo critério acima, serve como uma

indicação do potencial de reutilização dos dados. Embora os princípios FAIR indiquem que os dados reutilizáveis "atendem aos padrões da comunidade relevantes para o domínio" (WILKINSON *et al.*, 2016), as plataformas que são capazes de atender às necessidades dos pesquisadores em domínios correlatos têm o potencial de facilitar a pesquisa. Os dados devem "usar uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento" para serem considerados interoperáveis (WILKINSON *et al.*, 2016). Assim, os repositórios que possuem um esquema específico (fixo/relacional) o utilizam como uma forma de facilitar essa interoperabilidade; no entanto, esquemas fixos podem limitar os pesquisadores em suas decisões sobre quais dados eles podem ou não armazenar. Finalmente, a legibilidade dos dados por máquina foi uma ênfase dos proponentes dos princípios FAIR. Como tal, o que estamos chamando "*lifting* de dados" fornece mecanismos inerentes para cumprir cada um desses princípios.

As seções a seguir detalham as características específicas de cada um dos repositórios dos três *data centers* pesquisados, com relação aos quatro critérios de avaliação acima. O grau de atendimento aos critérios pode ser considerado uma proxy para se entender como os princípios FAIR são mais amplamente considerados pelos *data centers*.

5. Análise dos repositórios pesquisados

Segundo BYRD *et al.* (2020), sempre que for viável, os dados de pesquisa científica devem ser compartilhados por meio de repositórios específicos de domínio, que usam tipo de dados amplamente empregados em um campo. Tais repositórios específicos são armazéns de dados (*warehouses*) ideais. Eles fornecem acesso de longo prazo aos dados por meio do fornecimento de IDs persistentes, como os identificadores de objetos digitais (DOI). Eles reduzem os custos de pesquisa ao disponibilizar grandes coleções de dados correlatos em um local único, podendo reduzir o trabalho redundante e encorajar a geração de novas hipóteses a partir de análises secundárias. Por último, eles permitem que os dados sejam citados, fazendo com que os cientistas que geram dados acumulem crédito por compartilhar conjuntos de dados. Analisamos a seguir dois repositórios específicos de domínio.

5.1 Repositórios específicos de domínio

5.1.1. ImmPort

O ImmPort[3] é financiado pelo NIH com foco em "Bioinformática para o futuro da Imunologia". É um "portal de curadoria e distribuição" cujo objetivo é promover

3 https://www.immport.org/

o compartilhamento de dados imunológicos (BHATTACHARYA *et al*., 2018); trata-se de "um dos maiores repositórios abertos e com curadoria" de dados imunológicos humanos (SANSONE; CRUSE; THORLEY, 2017). Em seus esforços de curadoria de seus dados, o ImmPort elabora diretrizes e padrões com base em sugestões da comunidade de pesquisa em imunologia, maximizando a acessibilidade e interoperabilidade dos dados desta comunidade. O repositório é composto por quatro componentes: **dados privados**, **dados compartilhados**, **análise de dados** e **recursos**. Os dados coletados são selecionados no componente de dados privados e, eventualmente, publicados por meio do componente de dados compartilhados. O componente de análise de dados utiliza a ferramenta *Galaxy* para permitir a análise de dados no espaço do próprio repositório. O Galaxy facilita a análise e meta-análise de dados de citometria, foco do portal. Finalmente, em **recursos** reúne-se informações sobre ImmPort, suas publicações e tutoriais.

Diferentemente de outros repositórios, o ImmPort usa ontologias como uma forma de anotar seus dados com termos comuns e consensuados, incluindo uma ontologia Celular, uma de Doenças, uma para Investigações Biomédicas, uma de Proteínas e uma ontologia de Vacinas. Essas ontologias foram utilizadas na elaboração do *ImmPort Data Model*, que detalha as variáveis armazenadas em cada tabela e as relações entre elas. Ao fazer *upload* de dados para o ImmPort, o modelo de dados fornece um conjunto de termos comuns a serem usados para que a anotação seja consistente com os demais dados já adicionados ao repositório. Isso é aplicado por meio do uso de modelos de *upload* de dados e de uma ferramenta de validação. Os estudos disponíveis no repositório podem ser consultados por meio de uma busca básica por palavra-chave ou pela aplicação de filtros que incluem metadados como: se o estudo foi ou não um ensaio clínico, o tipo do estudo, o foco da pesquisa, as espécies pesquisadas, o tipo de amostra biológica e o tipo de ensaio clínico.

Esses recursos de pesquisa não consideram os dados em si, apenas certos metadados fornecidos no momento da submissão do estudo ao repositório, o que é feito por meio de templates pré-definidos. Assim, não há como consultar dados filtrando certos critérios em vários estudos simultaneamente. Além disso, para visualizar os dados em si, os arquivos individuais devem ser baixados pelo pesquisador. No entanto, metadados detalhados sobre os estudos armazenados estão disponíveis diretamente no site. Os metadados são padronizados por meio de templates elaborados com base no modelo de dados.

O domínio do ImmPort é estritamente o da imunologia. Percebe-se que a plataforma foi adaptada especificamente para acolher pesquisas orientadas à imunologia. O próprio modelo de dados também possui elementos muito específicos da imunologia. Embora essa rigidez seja importante quando os dados se ajustam ao

modelo, ela limita a utilização do ImmPort por outras pesquisas de domínios que poderiam cruzar com a imunologia. O ImmPort também tem um esquema específico por meio do modelo de dados e, portanto, claramente não é livre de esquema (*schema-free*). Isso limita quando os pesquisadores precisam armazenar dados que não se enquadram exatamente no esquema.

O ImmPort tem algum nível de *lifting* de dados, embora não esteja claro exatamente o quão significativo e abrangente ele seja. O repositório fornece aos pesquisadores modelos para usar ao formatar e enviar seus dados e pede que os pesquisadores validem esses dados em relação aos modelos existentes. Isso mostra que o ImmPort visa padronizar seus dados para que os estudos sejam compatíveis entre si. No entanto, como os estudos são baixados arquivo por arquivo, não está claro se portal usa ou não esses modelos para armazená-los de uma forma a serem combinados, gerando informações e conhecimentos. No geral, ImmPort mostra certos recursos interessantes, mas não ficou claro o quão profundamente ele aplica o processo de *lifting* para armazenar seus dados.

Considerando *data centers* específicos de domínio financiados pelo NIH, além o ImmPort, o *National Data Archive* é uma infraestrutura para hospedar repositórios de dados no domínio da saúde mental.

5.1.2. NIMH National Data Archive (NDA)

Inicialmente desenvolvida para integrar um conjunto de repositórios de dados de pesquisa como o *National Database for Autism Research* (NDAR[4]) e outros três em saúde mental, "se tornou uma plataforma para compartilhar dados em saúde mental e outras pesquisas". A plataforma tem restrições estritas de uso de dados, e o *download* requer que o usuário preencha uma Certificação de Uso assinada pelo NIH. Embora isso limite o uso da plataforma, resumos dos dados estão disponíveis e podem ser consultados. O NDA se ramificou para incluir outros aspectos do domínio da saúde mental. Os repositórios incluídos, além do NDAR, são o *Research Domain Criteria Database* (RDoCdb), o *National Database for Clinical Trials related to Mental Health* (NDCT) e o NIH MRI *Repository* (PedsMRI). O NDA está estruturado para atender às necessidades de dados específicos da pesquisa em saúde mental. Além disso, a restrição de seu acesso o torna acessível predominantemente aos integrantes das comunidades da área de saúde mental.

O conteúdo do NDA é organizado em torno do conceito de "*Identificador Único Global*" (GUID), que serve como uma forma de identificar dados de indivíduos únicos (DAN *et al.,* 2018). Os GUIDs são gerados por uma ferramenta que exige que o

4 https://nda.nih.gov/

pesquisador insira informações de identificação pessoal específicas, que são então usadas para gerar um código *hash* que representa unicamente o conjunto de dados. As mesmas informações de identificação pessoal garantirão que o mesmo GUID seja gerado, portanto, se o mesmo sujeito participar de vários estudos diferentes, ele não será duplicado no sistema. Isso permite que todos os dados sejam internamente relativos a uma única pessoa, possibilitando ao NDA fornecer consultas sofisticadas para extração de dados.

O NDA tem seis ferramentas de consulta: consulta geral, dados de laboratórios (*Data from Labs*), dados de artigos (*Data from Papers*), dicionário de dados, consulta por conceito e consulta por GUID. Cada ferramenta fornece seus próprios recursos exclusivos, o que aprimora o processo de coleta e análise de dados. A consulta geral permite que o pesquisador selecione campos predefinidos para construir uma consulta. Os resultados desta consulta são exibidos (junto com as estatísticas de resumo) e os dados resultantes podem ser baixados. Além disso, o usuário pode selecionar quais campos exatos deseja baixar, bem como de quais fontes. Isso é exclusivo porque significa que uma única consulta pode gerar resultados em todos os repositórios que usam NDA para armazenar seus dados (embora sejam necessárias certificações de uso de dados para baixar os dados de cada repositório). As ferramentas *Data from Labs* e *Data from Papers* consultam informações de coleções de NDA e estudos de NDA, respectivamente. Aqui, coleções e estudos podem ser filtrados por diferentes critérios e baixados usando o mesmo mecanismo de *download* usado pela ferramenta Consulta Geral. Isso é crucial porque permite que um pesquisador selecione várias coleções ou estudos e extraia estruturas específicas, conforme definidas no Dicionário de Dados, apenas para *download*. A ferramenta *Data Dictionary* permite ao pesquisador selecionar "estruturas de dados" e "elementos de dados" diretamente do dicionário de dados. Este dicionário mostra os vários atributos de cada estrutura de dados e inclui informações detalhadas sobre os elementos que contém. Finalmente, a ferramenta *Query by Concept* permite consulta por meio de "conceitos ontológicos", conforme definido pela ASD *Phenotype Ontology*, usando os mesmos recursos de filtragem e *download* disponíveis na plataforma como um todo. É importante notar, entretanto, que os dados não são armazenados usando nenhum tipo de representação ontológica, a ontologia é usada apenas como uma ferramenta de filtragem. Na verdade, a abordagem do NDA "não permite a criação fácil de uma ontologia, seja entre todos os dados nas avaliações clínicas no NDAR ou entre os dados no NDAR e outros léxicos" (DAN *et al.,* 2018).

Assim como o ImmPort, o NDA usa um esquema muito específico, conforme definido em seu dicionário de dados. Quando um usuário submete qualquer conjunto de dados, ele deve passar na validação do dicionário de dados, caso contrário,

não será aceito no sistema. Esta ferramenta de validação está publicamente disponível para pesquisadores e irá alertar o pesquisador sobre os erros em seus dados para que possam ser corrigidos. Além disso, todos os conjuntos de dados devem ter um GUID, o que os restringem a serem relacionados a um único sujeito (isso faz sentido para dados clínicos e de saúde mental, mas torna a extensibilidade baixa entre domínios).

Caso o pesquisador necessite de uma estrutura não definida pelo dicionário de dados, ele pode enviar novas definições ao *Help Desk* do NDA para eventual implementação. Isso faz com que mesmo que a plataforma tenha um esquema muito específico, tal esquema é de certa forma aberto a modificações e acréscimos. Contudo, isso faz com que as alterações no esquema demorem para serem implementadas, pois toda a manutenção é feita manualmente pelos funcionários do NDA. Isso também se aplica ao *upload* de dados, que geralmente leva 4 meses para ser disponibilizado publicamente na plataforma. Até esse ponto, os dados permanecem em um estado privado para que os funcionários do NDA possam revisar e garantir a qualidade dos mesmos.

Devido ao esquema rígido e às ferramentas de validação do NDA, ele pode realizar o processo de ingestão de dados rapidamente. As ferramentas de consulta disponíveis indicam que os dados armazenados no NDA são transformados de seu estado original de *upload* para um formato em que todos os dados são armazenados, em torno do GUID. Isso permite a extração de conhecimento entre estudos, coleções e repositórios de uma forma que muitas outras plataformas não são capazes.

A capacidade de manipular seletivamente os dados para *download* cria muitas oportunidades para análises exclusivas de dados. A nossa investigação não foi capaz de esclarecer exatamente como esses dados são armazenados nos "bastidores", mas ficou claro que todos os dados estão associados a um único GUID em toda a plataforma de uma forma que pode ser facilmente consultada. Isso diferencia o NDA de muitos outros repositórios; no entanto, ainda há espaço para melhoria quando se trata de automação de *upload* e curadoria de dados.

5.2. Repositórios de arquivamento de uso geral

Os dois *data centers* examinados até aqui são específicos de domínio, contudo, em certas circunstâncias, particularmente no início do desenvolvimento de um domínio científico de dados, pode não haver repositórios específicos. Nesses casos, os investigadores ainda podem escolher colocar os dados em plataformas de arquivamento de uso geral, como Figshare ou Zenodo, junto com metadados que descrevam precisamente os arquivos incluídos e seu formato. Para dados que não podem ser compartilhados publicamente devido a questões de privacidade, a pla-

taforma Synapse fornece uma plataforma de arquivamento de uso geral semelhante que suporta compartilhamento de acesso controlado (BYRD *et al.* 2020).

5.2.1. A Plataforma SYNAPSE

A Synapse[5] é uma plataforma de software de código aberto destinada aos pesquisadores que podem utilizá-la como local para armazenar e anotar seus dados. Ao contrário do ImmPort, ela não tem requisitos para a formatação dos dados, servindo apenas pesquisadores que desejam armazenar seus dados em algum lugar. Mesmo assim, vários *data centers* financiados pelo NIH a utilizam como repositório, e a própria plataforma é financiada por vários institutos ligados ao NIH.

A plataforma permite que seus usuários criem espaços de trabalho pessoais, carreguem arquivos diferentes, conectem-nos por meio de relações de proveniência, anotem arquivos para melhor descoberta, forneçam uma narrativa para os dados, criem identificadores de objetos digitais (DOI) para qualquer recurso e trabalhem colaborativamente. Para se registrar como usuário no Synapse basta fornecer um e-mail, e o *download* de dados públicos e a criação de conteúdo ficam facilmente acessíveis. A ressalva é que para armazenar dados sobre seres humanos (uma vez que existem restrições de uso específicas para isso), deve-se passar por um processo de certificação.

O Synapse pode ser operado por meio de vários métodos, incluindo Python e R, além da interface web tradicional. No entanto, certas funcionalidades (como baixar um grupo de arquivos) estão disponíveis apenas via Python ou por linha de comando. Cada recurso na plataforma tem um SynapseID exclusivo e, portanto, pode ser recuperado. O usuário deve usar a interface web para determinar o SynapseID do recurso, mas uma vez encontrado, ferramentas automatizadas podem conduzir a análise de dados.

Cada projeto no Synapse armazena suas informações em tabelas relacionais cujos esquemas são definidos pelo proprietário do projeto. Isso faz com que um projeto possa ser consultado usando comandos semelhantes aos da linguagem SQL, mas uma interface de pesquisa / filtro padrão também está disponível. Geralmente, essas consultas são utilizadas para pesquisar arquivos específicos ou múltiplos arquivos que compartilham uma característica específica, porém o usuário deve conhecer o esquema para gerar uma consulta bem-sucedida. O Synapse suporta duas estruturas diferentes: visualizações de tabelas e visualizações de arquivos. As visualizações de arquivo permitem navegar pelos arquivos carregados, visualizá-los e baixá-los. Contudo, as consultas não podem ser executadas nos próprios da-

5 https://www.synapse.org/

dos. Já uma tabela de dados pode ser pesquisada e consultada. As consultas podem se estender por diferentes repositórios, já que cada recurso hospedado no Synapse tem um ID exclusivo; no entanto, como os esquemas de tabela são identificados pelo proprietário do projeto, não há garantia de que uma única consulta possa se estender com êxito por vários repositórios que usam esquemas diferentes.

Como o Synapse opera apenas como uma plataforma independente de domínio, não há um domínio específico conectando todos os repositórios. Qualquer usuário certificado pode fazer *upload* e armazenar dados na plataforma, um pesquisador não precisa operar em nenhum domínio específico para ter os benefícios do site. A Synapse é *schema-free* no sentido de que cabe ao usuário decidir que esquema usar para armazenar os dados. No entanto, cada recurso em um projeto deve seguir um esquema padrão pré-definido para que o projeto seja carregado e consultável. Se o esquema não for suficiente para um pesquisador, ele pode criar seu projeto com seu próprio esquema, mas isso resulta na separação de um repositório existente do qual eles poderiam se beneficiar.

Tabelas relacionais são primitivas quando se trata de *lifting* de dados, pois suas estruturas rígidas limitam as informações que podem ser extraídas. Como tal, o Synapse é limitado quando se trata de seus recursos de *lifting* de dados. Isso pode ser percebido pela dificuldade da plataforma em consultar estudos e seus recursos. No entanto, o Synapse tem uma capacidade interessante que permite rastrear a proveniência dos dados. Este sistema de proveniência permite aos usuários "rastrear o histórico de análise e comunicar e compartilhar uma sequência de etapas de processamento". O próprio utilizador deve definir a proveniência (de preferência quando carrega ou edita um arquivo), caso contrário, perde-se o rastro da proveniência. No geral, a abertura do Synapse e a sua baixa barreira de entrada tornam a plataforma amplamente acessível, mas essa liberdade vem ao custo de impossibilitar abordagens padronizadas de troca de dados entre repositórios.

6. Conclusão

Obviamente, a organização e disponibilidade de dados para pesquisas em saúde estão altamente relacionadas à quantidade de informações e conhecimentos que pode ser extraído deles. É por isso que grupos de pesquisa estão começando a desenvolver e aplicar ferramentas para melhor estruturar esses dados para que possam ser analisados e compartilhados mais prontamente. Em muitos casos, esses grupos recorrem aos princípios orientadores FAIR como uma referência para desenvolver os seus recursos funcionais de compartilhamento de dados. Disponibilizar dados dessa forma, seguindo tais princípios, diminui as barreiras inerentes à realização bem-sucedida de pesquisas em saúde, permitindo que os pesquisadores utilizem

e apliquem em suas pesquisas, dados que muitas vezes não foram coletados por eles próprios. Isso também abre portas para estudos cruzados e colaborativos entre domínios, uma tendência que tem aumentado constantemente nos últimos anos.

Dos atualmente 99 repositórios financiados pelo NIH, alguns têm recursos que ilustram o movimento em direção ao maior compartilhamento de dados científicos por pesquisadores, especialmente quando se trata de *lifting* de dados, que é fundamento importante para que os dados sejam considerados FAIR. Todos os três repositórios analisados apresentam certos atributos que mostram seu movimento em direção ao uso mais profundo do *lifting* de dados em sua infraestrutura, embora ainda sejam bastante limitados na adoção mais ampla deste critério.

Conforme sintetizado na Tabela 1, esses repositórios foram avaliados por meio de quatro critérios - estrutura de consulta, domínio científico de atuação, esquema de representação dos dados e *lifting* de dados - como uma forma de entender o grau de sua sofisticação no caminho para atender aos critérios de dados FAIR e legíveis por máquina. Embora cada um dos repositórios analisados tenha seus pontos fortes e fracos, é importante entender o que estas organizações estão fazendo para aprimorar seus recursos voltados ao compartilhamento de dados para pesquisas futuras, como uma forma de compreender o estado da informação e do conhecimento na pesquisa em saúde como um todo.

Tabela 1 - Síntese analítico-comparativa dos repositórios avaliados.

Repositórios/ Critérios	ImmPort	NDA	Synapse
estrutura de consulta	Por termos e filtros por metadados	Por dados laboratoriais, por dados de artigos, por dicionário de dados, por conceito e por GUID	Por tabela e por arquivo
domínio científico	Imunologia	Saúde mental	qualquer
esquema de dados	específico do repositório	específico do repositório	*específico do projeto*
lifting de dados	limitado	limitado	limitado

Fonte: Elaborada pelo autor.

7. Referências

BHATTACHARYA, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data* 5**,** 180015 (2018). https://doi.org/10.1038/sdata.2018.15

BYRD, James Brian *et al*. "Responsible, practical genomic data sharing that accelerates research." **Nature Reviews Genetics** 21.10 (2020): 615-629.

DAN, Hall *et al.* Sharing Heterogeneous Data: The National Database for Autism

Research. **Neuroinformatics** 10.4 (2012): 331–339. PMC. Rede. 2 de maio de 2018.

MCQUILTON, Peter *et al.* BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences, **Database**, Volume 2016, 1 January 2016.

PAN, Jeff Z. *et al.* Exploiting Linked Data and Knowledge Graphs in Large Organizations. Springer International Publishing. 2017.

PEREZ-RIVEROL, Yasset *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. **Nature biotechnology** 35.5 (2017): 406.

PIÑERO, Janet *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, **Nucleic Acids Research**, Volume 45, Issue D1, 4 January 2017, Pages D833–D839.

SANSONE, Susanna-Assunta; CRUSE, Patricia; THORLEY, Mark. High-quality science requires high-quality open data infrastructure. **Sci. Data** 5:180027 doi: 10.1038/sdata.2017.27 (2017).

VITA, Randi *et al.* FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. **Database**, Volume 2018, 1 January 2018.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v3.1, p. 1-9. 2016.

Como citar com o DOI individual

BAX, Marcello Peixoto. Investigando os princípios FAIR em repositórios de dados científicos do National Institutes of Health (NIH). *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 171 - 186. DOI: 10.22477/9786589167242.cap13

Seção 4

DADOS REUSÁVEIS

Reúso de dados: princípios FAIR e o ecossistema de pesquisa

Sônia Elisa Caregnato[1], Rafael Port da Rocha[2], Rene Faustino Gabriel Junior[3]

1. Introdução

O movimento da ciência aberta ganhou impulso nos últimos anos e isso decorre tanto das possibilidades tecnológicas que se concretizaram como da percepção pela sociedade de que a pesquisa científica é uma atividade coletiva, financiada publicamente e que precisa retornar valor para a sociedade que a apoia.

Nesse sentido, os dados produzidos no curso de uma pesquisa, além das publicações que os contextualizam, devem estar em acesso aberto para que possam ser compartilhados entre os cientistas e reutilizados em novas pesquisas, retroalimentando a ciência, cujo caráter é cumulativo. O compartilhamento de dados, portanto, tem sido uma demanda de governos, de agências financiadoras e de instituições de pesquisa, mas para que isso possa se concretizar, são necessárias ações de planejamento, gestão e curadoria do conjunto de dados em repositórios. Essas ações ocorrem no escopo de um ecossistema de pesquisa que envolve tecnologias, pessoas e instituições.

Com a finalidade de assegurar boas práticas para o compartilhamento de dados de pesquisa, por meio dos princípios FAIR se estabelece que os dados devam ser fáceis de encontrar (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e reutilizáveis (*Reusable*). A intenção é, por intermédio dos princípios,

1 Doutora pela University of Sheffield, Reino Unido. Professora dos Programas de Pós-Graduação em Comunicação e em Ciência da Informação, ambos da Universidade Federal do Rio Grande do Sul (UFRGS). E-mail: sonia.caregnato@ufrgs.br

2 Doutor em Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professor do Programa de Pós-Graduação em Ciência da Informação da UFRGS. E-mail: rafael.rocha@ufrgs.br

3 Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp). Professor do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul (UFRGS). E-mail: rene.gabriel@ufrgs.br

facilitar a reutilização dos dados, tanto por humanos como pelas máquinas (wilkinson et al., 2016).

Compartilhamento e reúso, portanto, representam um par de conceitos que se complementam. No entanto, como observado por alguns autores (pasquetto et al., 2017; tenopir et al., 2011; wallis et al., 2013), o primeiro é muito mais frequentemente estudado do que o segundo, embora os benefícios do compartilhamento somente possam ser obtidos se os dados forem efetivamente reutilizados.

Em busca de uma melhor compreensão sobre esse tema e suas implicações, este trabalho explora a utilização do termo *reúso de dados de pesquisa*, assim como de termos comumente relacionados a este, por meio de análise da literatura nacional e internacional. Inicialmente, aborda-se o compartilhamento de dados de pesquisa, para depois relacioná-lo ao reúso e aos princípios fair. Na sequência, discute-se o significado dos termos *uso* e *reúso*, bem como suas variações para, finalmente, discorrer sobre as condições necessárias para uma efetiva reutilização de dados de pesquisa.

2. Compartilhamento de dados de pesquisa

O compartilhamento de dados de pesquisa está intrinsecamente relacionado ao seu reúso, seja para validá-los ou para dar origem a novas interpretações. No entanto, o ato do compartilhamento informa muito pouco acerca do uso que será feito dos dados. Há, sim, um pressuposto nas políticas de acesso aberto, qual seja, de que dados de pesquisa são úteis para outros pesquisadores e que eles vão reutilizá-los (pasquetto et al., 2019).

Segundo Borgman (2012), o compartilhamento diz respeito ao ato de abrir dados de forma que possam ser reutilizados por outros indivíduos. É importante destacar que o grau de confiança, a utilidade e o valor desses dados variam enormemente: enquanto alguns são estruturados e recebem curadoria, outros são simplesmente disponibilizados.

Em parte, pelo menos, a qualidade dos dados disponibilizados depende da forma como o compartilhamento é feito. Pasquetto et al. (2017) esclarecem que isso pode ser feito por meio de trocas privadas entre pesquisadores, depósito em repositórios, disponibilização em *websites* de laboratórios, suplemento de artigos de periódicos e, mais recentemente, artigos de dados, que esclarecem a proveniência e possibilitam a alocação de crédito aos autores. Para as autoras, assim sendo, tanto as trocas diretas entre pesquisadores como a disponibilização aberta de dados são entendidas como compartilhamento. O modo como isso é feito, segundo elas, varia conforme a área do conhecimento, os tipos de dados, o país, a agência de fomento ou outros fatores.

Boté e Térmens (2019), em contrapartida, preferem diferenciar o compartilhamento privado do compartilhamento público em repositórios institucionais, gerais ou especializados, chamando o segundo de publicação. Para eles, o compartilhamento entre pares durante o período em que os dados não estão necessariamente disponíveis publicamente pressupõe um nível grande de confiança entre as partes, bem como maior facilidade para se obter informações por meio de canais informais sobre como integrar esses dados em novos projetos. Publicações de dados em repositórios, no entanto, demandam um esforço maior visto a necessidade de providenciar documentação detalhada que acompanhe os dados e explique como usá-los.

Independentemente da forma, há, evidentemente, vantagens e benefícios reconhecidos no compartilhamento dos dados. Borgman (2012) identificou quatro razões para essa prática, quais sejam: a) reproduzir ou verificar resultados; b) tornar disponível os resultados da pesquisa financiada pelo poder público; c) possibilitar que outros formulem novas questões de pesquisa com base nos dados existentes; d) avançar o estado da pesquisa e da inovação.

No entanto, há também barreiras e desafios ao compartilhamento dos dados. Tanto é que, por ser um fenômeno intrincado e complexo, a professora Christine Borgman o chamou de dilema (*conundrum*) em seu célebre artigo de 2012. Segundo ela,

> Para que as recompensas da inundação de dados sejam colhidas, os pesquisadores que produzem esses dados devem compartilhá-los e fazê-lo de forma que os dados sejam interpretáveis e reutilizáveis por outros. Subjacente a esta declaração simples, estão camadas espessas de complexidade sobre a natureza dos dados, pesquisa, inovação e bolsa de estudos, incentivos e recompensas, economia e propriedade intelectual e políticas públicas. (BORGMAN, 2012, p. 1.059, tradução nossa).

Entre as dificuldades para o não compartilhamento dos dados de pesquisa estão o receio de uso inadequado dos dados e da competição, o custo para preparar os dados e a documentação, a falta de tempo, a ausência de uma infraestrutura e de padrões apropriados, questões éticas – entre elas, o uso dos dados para finalidades distintas daquelas para as quais foram coletadas – e, não menos importante, o fato de que os dados brutos são pouco úteis para a reutilização sem um esforço significativo da parte de quem os disponibiliza, para torná-los passíveis de novas análises (BORGMAN, 2012; KIM; YOON, 2017; PERRIER et al., 2020; ROWLEY et al., 2017; TENOPIR et al., 2011; WALLIS et al., 2013).

A defesa do compartilhamento de dados de pesquisa e a luta pela superação de obstáculos à sua reutilização são motivadores da busca por diretrizes que orientem as ações de gestão necessárias. Os princípios FAIR são um marco importante nesse esforço e também na ampliação do valor dos dados para seu reúso, conforme será abordado a seguir.

3. Princípios FAIR e o ecossistema de pesquisa

Além de ser um dos quatro princípios FAIR, o reúso é, ainda, a finalidade dos processos de curadoria de dados. Como a principal característica dos princípios FAIR é a oferta de um conjunto de orientações concisas, de alto nível, que valem para qualquer domínio e que devem ser aplicadas não somente aos dados, mas também aos metadados, aos identificadores, ao *software* e aos planos de gestão de dados, então eles se apresentam como facilitadores e orientadores do reúso.

Em relatório final intitulado Turning FAIR into reality (COLLINS et al., 2018), o Grupo de Especialistas em dados FAIR da Comissão Europeia apontou que a implementação dos princípios exige a criação de uma nova cultura de pesquisa, além de um ecossistema técnico constituído de serviços e infraestrutura apropriados, que incluem políticas, planos de gestão de dados, identificadores persistentes, padrões de interoperabilidade, metadados e repositórios. Além disso, os autores salientaram que é necessária a promoção do desenvolvimento de competências para, de um lado, processar e analisar dados (ciência de dados), e de outro, gerenciá-los e preservá-los ao longo de seu ciclo de vida (curadoria de dados), bem como desenvolver métricas para avaliar a conformidade aos princípios e, finalmente, buscar pela sustentabilidade e financiamento dos projetos.

O relatório define o ecossistema FAIR como um modelo que indica os componentes mínimos necessários para promover a criação, a curadoria e o reúso de objetos digitais FAIR de forma efetiva e sustentável (COLLINS et al., 2018). O elemento central desse ecossistema, portanto, é o objeto digital (os dados e outros recursos de pesquisa), que precisa estar acompanhado de identificadores persistentes e metadados para possibilitar que seja encontrado, usado e citado. Os objetos digitais FAIR precisam também estar representados em formatos de arquivos idealmente abertos e utilizar vocabulários comuns às comunidades, para que seja possível a interoperabilidade e o reúso. Além disso, a documentação associada deve incluir instruções acionáveis por máquina sobre as condições de uso e licenças permitidas. Por fim, como salientam Koers et al. (2020), esse ecossistema é sustentado por métricas, mecanismos de certificação, incentivos, financiamento e treinamento.

Um movimento importante para que os princípios FAIR se efetivem é a iniciativa GO FAIR, surgida na Europa, em 2017, que se expandiu para outros países,

inclusive o Brasil. O GO FAIR propõe a formação de redes de implementação de dados e serviços FAIR, de modo que os interessados possam trabalhar de forma participativa e colaborativa (SALES et al., 2020).

Nem os princípios FAIR nem a iniciativa GO FAIR estabelecem a obrigatoriedade de todos os dados de pesquisa serem abertos, ou seja, dados podem ser FAIR e, ao mesmo tempo, serem compartilhados de forma restrita. Essa provisão é necessária em determinadas circunstâncias, por exemplo, quando os dados incluem informações pessoais, confidenciais ou de valor comercial. Os maiores benefícios para a ciência e para a sociedade, no entanto, ocorrem quando os dados são tanto FAIR como abertos, pois a ausência de restrições aumenta as possibilidades de reúso em grande escala (COLLINS et al., 2018), ou, como afirmam Henning et al. (2019, p. 394), "[...] quanto mais abertos estiverem, mais serão usados, reusados e combinados com outros dados, promovendo o crescimento econômico, a inovação e o desenvolvimento".

As informações sobre licenças de uso, no entanto, devem ser claramente especificadas para que os dados sejam considerados FAIR. Assim, se os dados não puderem ser abertos, ou se somente puderem ser usados com restrições, essas informações deverão ser explicitadas. Em relação à reutilização, os princípios FAIR estabelecem que (WILKINSON et al., 2016, tradução nossa):

> 1. Os (meta)dados são detalhadamente descritos com uma pluralidade de atributos precisos e relevantes;
> 1.1 Os (meta)dados são publicados com licenças de uso de dados claras e acessíveis;
> 1.2 Os (meta)dados são associados a informações detalhadas sobre sua proveniência;
> 1.3 Os (meta)dados atendem a padrões relevantes para a comunidade da área.
> Ou seja, para serem reutilizáveis, tanto os dados como os metadados devem estar acompanhados de informações que efetivamente lhes possibilitem ser empregados em contextos diferentes daqueles em que foram criados.

Nesse sentido, o relatório Turning FAIR into reality (COLLINS et al., 2018) sugere 27 recomendações, cada uma acompanhada de um conjunto de ações relevantes para apoiar a efetivação do ecossistema de dados FAIR. Entre estas, há algumas que são diretamente relacionadas ao reúso. Particularmente, os autores recomendam que os financiadores de pesquisa deveriam incentivar o reúso de dados FAIR requerendo que as comunidades reutilizem conteúdos já existentes, quando possível. Isso pode ser feito ao solicitar aos pesquisadores que demonstrem nos projetos

de pesquisa que dados FAIR foram buscados e/ou consultados antes de propor a criação de novos dados, ou então ao reconhecer que os resultados de pesquisas que reutilizaram dados têm o mesmo valor das pesquisas que criam novos conteúdos, ou ainda, ao financiar pesquisas que reutilizem dados FAIR.

Para que o reúso efetivamente derive do compartilhamento, é necessário compreender todas as dimensões do fenômeno, iniciando pela própria definição do termo. Assim, a próxima seção trata especificamente do reúso dos dados, considerando suas dimensões e características.

4. Reuso de dados de pesquisa

Como mencionado anteriormente, todo compartilhamento de dados de pesquisa pressupõe a sua reutilização para benefício da própria ciência, da comunidade científica e da sociedade em geral, ou seja, na perspectiva do ecossistema de pesquisa. Para isso, inicialmente é necessário esclarecer o significado de reúso nesse contexto específico e na sua relação com o uso de dados de pesquisa.

Segundo van de Sandt e colegas (2019), o termo *reúso de dados de pesquisa* se refere a um conceito complexo que varia de acordo com as áreas do conhecimento. Mesmo assim, dizem as autoras, é fundamental definir esse termo porque cada vez mais este vem sendo utilizado por instituições de financiamento e de pesquisa, o que demonstra sua importância.

A primeira distinção frequentemente encontrada é aquela entre uso e reúso, conforme sugerida por Pasquetto et al. (2017, 2019). Outros conceitos relacionados também são apontados na literatura, por exemplo, reprodutibilidade, replicabilidade, integração e reanálise (BOTÉ; TÉRMENS, 2019; CURTY, 2019; VAN DE SANDT, 2019). Esses aspectos, incluindo as propostas de taxonomia ou modelos para compreender os tipos de dados de pesquisa serão discutidos a seguir.

4.1. Definindo reúso

Comumente, nos textos sobre dados de pesquisa, reúso é definido como o uso subsequente, feito por outros pesquisadores, dos dados coletados para determinado projeto, ou como definem com precisão Boté e Térmens (2019, p. 329, tradução nossa), "[...] encontrar, processar e analisar os conjuntos de dados de outros para criar novos conhecimentos".

Três elementos são fundamentais nessa definição: a) trata-se de um uso secundário, não originalmente previsto; b) que acontece temporalmente depois do uso; e c) e é feito por pesquisador (grupo de pesquisa) diferente daquele que coletou os dados. Quanto ao primeiro e segundo aspectos, parece não haver divergências na literatura. Em relação ao terceiro, no entanto, há cisões. Pasquetto et al. (2017) pon-

tuam explicitamente que, se um mesmo cientista retorna para o mesmo conjunto de dados em projeto posterior, essa ação seria caracterizada como de uso e não de reúso. Para as autoras, reúso ocorre quando os conjuntos de dados são recuperados por terceiros e empregados em outro projeto. Alguns autores, porém, não estabelecem essa distinção, por exemplo, Custers e Uršič (2016), enquanto outros ainda defendem expressamente que qualquer uso subsequente, mesmo aquele feito por quem obteve os dados, deva ser considerado reúso (CURTY, 2019).

Uma posição pouco usual é aquela assumida por van de Sandt e colegas (2019), ao concluírem que as características do discurso da área não provam que há alguma diferença entre reúso e uso. Com base na análise etimológica das palavras *uso* e *reúso* e nos conceitos relacionados, bem como na análise de discurso e na formulação de cenários, as autoras consideram quatro características frequentemente utilizadas para diferenciar uso de reúso:

a) o caráter dos dados, que se refere à quantidade de *datasets* reutilizados ou à sua transformação pelo reúso;
b) o usuário, que a literatura diferencia do produtor dos dados;
c) o propósito, que está relacionado à questão de pesquisa e/ou ao método;
d) a dimensão temporal, em que se evidencia o uso original antes do segundo uso dos dados.

Com base nessas análises, as autoras afirmam: "Dessa forma, nós definimos (re)uso como o uso de qualquer recurso de pesquisa, independentemente de quando é utilizado, do seu propósito, das suas características e de seu usuário" (VAN DE SANDT et al., 2019, p. 14). Ainda, elas buscam creditar a confusão dos termos ao modelo linear centrado no artigo publicado, que seria menos dinâmico e complexo que o panorama de pesquisa atual.

A originalidade e o brilhantismo do trabalho está também em relacionar essa proposta de simplificação da linguagem ao engajamento no movimento da ciência aberta, já que, afirmam as autoras, os pesquisadores estariam mais propensos a publicar e documentar uma pesquisa de qualidade para um propósito (o uso) que já está consolidado, ao invés de outro (reúso) que não é claramente entendido (VAN DE SANDT et al., 2019). Uma análise mais detalhada do artigo, porém, poderia revelar que se trata de uma estratégia relacionada à percepção de que o reconhecimento dos pares é maior quando se trata de trabalho original (uso) em vez de secundário (reúso). Isso poderia se manifestar na preocupação das autoras sobre a avaliação do impacto da pesquisa por meio de citações, que também é mencionado no texto.

De qualquer forma, grande parte da literatura que aborda o tema não parece propensa a tal mudança, pelo menos no curto prazo. Assim sendo, entende-se que essa diferenciação continuará, principalmente por ser útil no contexto de privacidade e proteção dos dados, conforme será discutido mais adiante. Na sequência, abordam-se as taxonomias ou os modelos que procuram diferenciar as formas de reúso dos dados.

4.2. Tipologias de reúso

Pode-se entender o reúso como uma categoria mais ampla, que engloba a reprodutibilidade, a replicabilidade, o reaproveitamento, a integração e a reanálise, entre outros termos – ou, então, com que este se relaciona (curty, 2019; pasquetto et al., 2017; van de sandt et al., 2019).

Partindo do contexto de *big data* e não dos dados de pesquisa, mas sem excluí-los, após identificar barreiras práticas, tecnológicas e legais, Custers e Uršič (2016) propõem uma taxonomia de reúso composta de três elementos: reciclagem (*recycling*), reaproveitamento (*repurposing*) e recontextualização (*recontextualization*). Reciclagem de dados diz respeito ao uso realizado diversas vezes, mas sempre com o mesmo objetivo inicial; reaproveitamento de dados refere-se ao reúso para propósitos distintos daquele para os quais eles foram primeiramente coletados; e recontextualização de dados implica a utilização em contextos diferentes dos quais eles foram inicialmente obtidos. Essa distinção é particularmente importante de uma perspectiva legal, em situações que envolvem privacidade e proteção dos dados pessoais, visto que normalmente a autorização dos sujeitos das pesquisas para a utilização dos dados é dada sob a condição de que o seu uso seja feito somente no âmbito daquele estudo.

Pasquetto et al. (2017), como base na literatura prévia, diferenciam reprodutibilidade de replicabilidade, por um lado, e integração de reúso independente, de outro. Reprodutibilidade ocorre quando um problema de pesquisa é formulado novamente em cima dos mesmos dados e métodos, a fim de validar, verificar ou confirmar a pesquisa, ao passo que replicação implica o uso de novos dados para responder, com os mesmos métodos, a uma questão anterior. O reúso independente diz respeito a um agente externo que realiza o reúso; por exemplo, a reprodução de um estudo é um exemplo de reúso independente. Integração envolve o reúso de *datasets* combinados com outros dados, sejam eles resultado de pesquisas de outros ou de novas observações.

Como se observa, não são dimensões excludentes ou que podem ser facilmente diferenciadas entre si. Assim, em estudo posterior, Pasquetto et al. (2019) introduzem outro tipo de diferenciação: o reúso é um contínuo que varia de comparativo até integrativo. O reúso comparativo de dados, como o termo indica, envolve usar

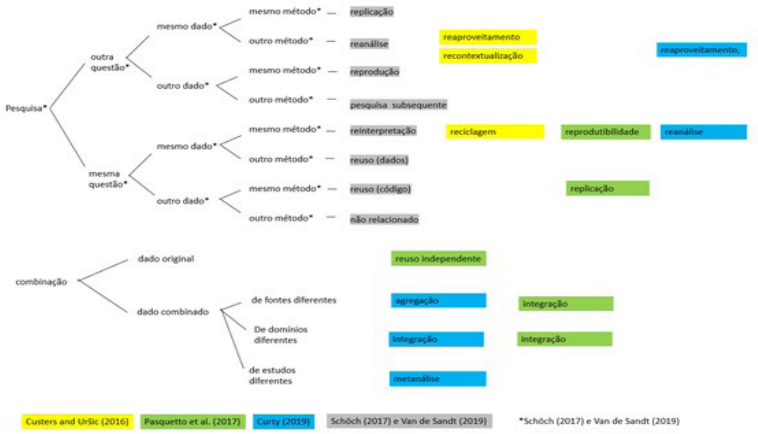
os dados para uma comparação específica, o que requer experiência em interação, ou seja, saber o suficiente sobre os dados para avaliar sua qualidade e valor. Já o reúso integrativo, como reutilizar dados em um novo experimento, envolve interpretação e, portanto, necessita de conhecimento científico mais especializado e aprofundado para ser realizado, assim como maior confiança na qualidade dos dados a serem reutilizados.

Curty (2019) propõe uma classificação que descreve cinco abordagens para reúso de dados de pesquisa, quais sejam, a) reaproveitamento, b) agregação, c) integração, d) metanálise e e) reanálise. Segundo a autora, no reaproveitamento, os dados de um único estudo são reutilizados integral ou parcialmente para novas análises, resultantes de questões de pesquisa diferentes, sem que sejam complementados ou integrados com dados de outras fontes. No reúso por agregação, reúnem-se dados provenientes de diferentes estudos/fontes para compor um *dataset* mais completo. O reúso por integração é aquele que combina dados de diferentes tipos de estudos, por meio de variáveis que ligam estudos separados. A metanálise combina dados provenientes de múltiplos estudos independentes, com perguntas de pesquisa muito semelhantes, integrando-os em uma análise mais ampla e substancial. A reanálise envolve a verificação dos resultados originais, por meio de nova análise, utilizando os mesmos métodos e técnicas, ou seja, trata-se do conceito de reprodutibilidade abordado por Pasquetto et al. (2017).

Como se observa, não há consenso entre os pesquisadores sobre uma definição para reúso nem mesmo sobre a forma de categorização de suas variáveis ou especificidades. Uma abordagem que parece promissora, por sistematizar os diferentes conceitos de reuso com base em três dimensões da pesquisa (questão, dado e método), é a de Schöch (2017). O autor deriva oito categorias de reprodutibilidade em pesquisa, sendo que duas destas (pesquisa subsequente e não relacionada) não apresentam proximidade com o reúso, diferentemente das seis outras, que estão diretamente relacionadas ao reúso, a saber: replicação, reanálise, reprodutibilidade, reinterpretação, reúso dos dados e reúso do código.

A fim de relacionar esses diferentes conceitos utilizados na literatura, propõe-se o agrupamento das classificações de Custers e Uršič (2016), Pasquetto et al. (2017), Curty (2019) e Schöch (2017) e van de Sandt (2019) (Figura 1). Esse agrupamento ocorre com base na observação de dois critérios gerais: tipos de reúso determinados pelo contexto da pesquisa e tipos de reúso determinados pela necessidade de combinação dos dados. Os tipos de reúso determinados pelo contexto da pesquisa são reagrupados com base nas categorias propostas por Schöch (2017) e van de Sandt (2019): mesmas/outras questões de pesquisa, mesmos/outros dados e mesmos/outros métodos de pesquisa.

Figura 1 – Categorização dos termos relacionados ao conceito de reúso de dados de pesquisa



Fonte: Os autores.

A Figura 1 também mostra que "reaproveitamento", "reciclagem", "recontextualização", de Custers e Uršič (2016), "reprodutibilidade" e "replicação", de Pasquetto et al. (2017), e "reaproveitamento" e "reanálise", de Curty (2019), observam o reúso de dados no contexto da pesquisa. Já a combinação do dado com outros é critério para "integração", "agregação" e "meta análise", de Curty (2019), e "reúso independente" e "integração", de Pasquetto et al. (2017). Curiosamente, em Schöch (2017) e van de Sandt (2019), há uma inversão entre termo e conceito, no que tange à "replicação" e "reprodutibilidade", com relação à Pasquetto et al. (2017).

Entende-se que a representação sintetiza as interpretações do termo *reúso* na literatura acerca dos dados de pesquisa, ao mesmo tempo que revela a complexidade das iniciativas de compreendê-lo efetivamente.

4.3. Condições para reúso de dados de pesquisa

Definições conceituais são fundamentais, mas o reúso efetivo dos dados de pesquisa somente poderá ocorrer se forem oferecidas condições aos pesquisadores e incentivadas as ações apropriadas no âmbito do ecossistema de pesquisa.

Em estudo conduzido para avaliar as práticas de reúso de dados de pesquisa em situações em que esse uso falhou, Yoon (2016) propôs formas de superação dos problemas. A autora oferece as seguintes sugestões:

a) a facilidade de reutilização, particularmente relacionada à interoperabilidade e ao acesso, é a condição inicial para experiências de sucesso com reúso de dados;

b) a compreensão dos dados por meio de documentação pode ser uma dificuldade menor, pelo menos para pesquisadores experientes, embora o processo ainda represente um desafio;
c) o principal componente da experiência de reúso que se torna falho é a falta de suporte na reutilização de dados, o que mostra a necessidade de desenvolver um sistema de apoio para quem reutiliza dados de pesquisa.

A importância da documentação é frequentemente enfatizada como condição fundamental para o sucesso do reúso. Por exemplo, Curty (2019) argumenta que os dados, para serem reutilizados, precisam ser considerados relevantes, completos, compreensíveis e confiáveis, e que esses atributos só podem ser observados se os dados estiverem acompanhados de informações suplementares e descrições sobre sua origem e processamento, ou seja, de documentação que os contextualizem. Muito embora não seja exatamente esse o resultado de Yoon (2016), a autora aponta para a necessidade de desenvolver competências nesse sentido, o que vai ao encontro do trabalho de Estevão e Strauhs (2020), que apontam para a exigência de letramento informacional no reúso de dados por parte dos pesquisadores, muitos dos quais não têm experiência no tema.

Os resultados do estudo de Kim e Yoo (2017) sobre o comportamento de reúso de dados de cientistas mostram que existem variações significativas entre as disciplinas, bem como dentro destas, nas intenções de reúso de dados. A utilidade dos dados, conforme é percebida pelos cientistas, a preocupação com a qualidade destes e a oferta de recursos nas suas organizações foram considerados os elementos mais importantes pelos entrevistados para o reúso de dados. No nível disciplinar, a disponibilidade de repositórios de dados mostrou uma significativa relação positiva com a intenção de reutilização de dados.

Em síntese, para que o reúso de dados de pesquisa se concretize e para que cumpra a promessa de aprimorar a forma e os resultados da produção do conhecimento em benefício de toda a sociedade, é necessário mobilizar todo o ecossistema de pesquisa: as pessoas necessitam de incentivos e capacitação; as instituições precisam fornecer as condições necessárias e as tecnologias devem ser exploradas no seu potencial total.

5. Considerações finais

O reúso de dados é a peça central no ecossistema de pesquisa, no qual os benefícios do compartilhamento dos dados somente serão efetivos se os dados forem preparados observando-se o reúso como princípio, como quando se adota os princípios FAIR. O reúso também é um dos objetivos da curadoria digital de dados

de pesquisa, pois esta compreende ações que manterão e agregarão valor a dados confiáveis para uso presente e futuro. Além disso, os valores desses dados somente poderão ser determinados por meio da compreensão adequada de seu reúso, no contexto da sua comunidade de usuários.

O termo *reúso de dados de pesquisa* remete a um conceito complexo, no sentido de auxiliar na compreensão do reúso por parte do pesquisador ou do curador de dados. Este trabalho buscou desenvolver a análise desse termo por meio de suas definições, de suas relações com outros termos e de características que promovem distinções entre uso e reúso. Também abordou as formas de reúso, por meio da identificação de seus tipos e categorias, e o reúso pela perspectiva do seu favorecimento.

A continuidade dos estudos sobre reúso é necessária, assim como a compreensão das perspectivas e práticas dos pesquisadores das diferentes comunidades científicas, pois o reúso de dados de pesquisa não será alcançado plenamente por meio da simples disponibilização dos dados em repositórios. Há um ecossistema que precisa ser mobilizado para possibilitar alcançar esse fim.

6. Referências

BERGHMANS, S.; *et al*. Open Data: the researcher perspective - survey and case studies, 2017. **Mendeley Data**, versão 1, 4 abr. 2017. DOI: 10.17632/bwrnfb4bvh.1

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**, v. 70, n.8, p.888–904, 2019.

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, p. 1059–1078, 2012. DOI:10.1002/asi.22634

BOTÉ, J.; TÉRMENS, M. Reusing Data: Technical and Ethical Challenges. DESIDOC **Journal of Library & Information Technology**, v. 39, n. 6, p. 329-337, 2019. DOI: 10.14429/djlit.39.6.14807

COLLINS, S.; et al. **Turning FAIR into reality**: Final report and action plan from the European Commission expert group on FAIR data. Bruxelas: European Commission, 2018. DOI:10.2777/1524.

CURTY, R. G. Abordagens de reúso e a questão da reusabilidade dos dados científicos. **Liinc em revista**, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4777

CUSTERS, B.; URŠIČ, H. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. **International Data Privacy Law**, v. 6, n. 1, p. 4–15, 2016. DOI: 10.1093/idpl/ipv028.

DIAS, G. A.; ANJOS, R. L. D.; ARAÚJO, D. G. A, Gestão dos dados de pesquisa

no âmbito da comunidade dos pesquisadores vinculados aos programas de pós-graduação brasileiros na área da ciência da informação: desvendando as práticas e percepções associadas ao uso e reúso de dados. **Liinc em revista**, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4683

ESTEVÃO, J. S. B.; STRAUHS, F. R. Letramento informacional para reúso de dados nas ciências sociais: requisitos e competências. **Informação & Informação**, v. 25, n. 2, p. 1-25, 2020. DOI: 10.5433/1981-8920.2020v25n2p1

HENNING, P. C.; et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, maio/ago. 2019. doi: http://dx.doi.org/10.19132/1808-5245252.389-412

KIM, Y.; YOON, A. Scientists' data reuse behaviors: a multilevel analysis. **Journal of the Association for Information Science and Technology**, v. 68, n.12, p.2709-2719, 2017. DOI: 10.1002/asi.23892

KOERS, H.; et al. Recommendations for Services in a FAIR Data Ecosystem. **Patterns**, v. 1, n. 5, 100058, 2020. DOI:10.1016/j.patter.2020.100058

PASQUETTO, I. V.; BORGMAN, C. L.; WOFFORD, M. F. Uses and Reuses of Scientific Data: The Data Creators' Advantage. **Harvard Data Science Review**, v.1, n.2, 2019. DOI: 10.1162/99608f92.fc14bf2d

PASQUETTO, I. V.; RANDLES, B. M.; BORGMAN, C. L. On the reuse of scientific data. **Data Science Journal**, v. 16, n. 8, 2017. DOI:10.5334/dsj-2017-008

PERRIER, L; BLONDAL, E.; MACDONALD, H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. **PLoS** ONE, v. 15, n. 2, e0229182, 2020. DOI:10.1371/journal.pone.0229182

SALES, L. et al. GO FAIR Brazil: a challenge for Brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. DOI:10.1162/dint_a_00046

SCHÖCH, C. Wiederholende Forschung in den digitalen Geisteswissenschaften. **Anais.** DHd2017: Digital Nachhaltigkeit (DHd2017), Bern, Switzerland, 13-18 February 2017. Bern: Universitat Bern, 2017. Disponível em: https://zenodo.org/record/277113#.X6D7u4hKi01 Acesso em: 14 out. 2020

TENOPIR, C. et al. Data sharing by scientists: practices and perceptions. PLoS ONE, San Francisco, v. 6, n. 6, e21101, 2011. DOI: 10.1371/journal.pone.0021101

VAN DE SANDT, T. et al. The definition of reuse. **Data Science Journal**, v. 18, n. 1, p. 1-19, 2019. DOI: https://doi.org/10.5334/dsj-2019-022

WALLIS, J.C; ROLANDO, E.; BORGMAN, C.L. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology**. PLoS** ONE, v. 8, n. 7, e67332, 2013. DOI: 10.1371/journal.

pone.0067332
WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, mar. 2016.
YOON, A. Red flags in data: Learning from failed data reuse experiences. **Proceedings of the Association for Information Science and Technology**, v. 53, n. 1, p. 1–6, 2016. DOI: https://doi.org/10.1002/pra2.2016.14505301126

Como citar com o DOI individual
CAREGNATO, Sônia Elisa; ROCHA, Rafael Port da; FAUSTINO, Rene; JUNIOR, Gabriel. Reúso de dados: princípios FAIR e o ecossistema de pesquisa. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). P**rincípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 187 - 200. DOI: 10.22477/9786589167242.cap14

#SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa

Silvana Aparecida Borsetti Gregorio Vidotti[1],
Emanuelle Torino[2], Caio Saraiva Coneglian[3]

1. Introdução

O momento atual da sociedade pode ser compreendido a partir de uma importante evolução das tecnologias digitais, mas especificamente de uma compreensão do papel dos dados no processo de tomada de decisão. Em linhas gerais, a coleta, o tratamento, a análise e a disponibilização de dados passaram a ser processos essenciais para que todos os setores da economia sobrevivessem e avançassem. Ademais, o próprio processo da pesquisa científica foi fortemente impactado por essa tendência, e passou a empregar de forma bastante significativa o uso, a análise e o reuso de dados para a realização de pesquisas.

Esse contexto levou ao desenvolvimento de estudos relacionados, por exemplo, à ciência de dados, ao *big data* e à pesquisa e comunicação científica, que possibilitaram compreender o papel dos dados no momento atual. Dessa forma, um aspecto importante para o entendimento e a análise do período que vivemos é o volume dos dados existentes que, quando tratados e analisados, são capazes de gerar uma grande riqueza para todos os processos envolvidos.

É essencial destacar nesse contexto o movimento de dados abertos, que está vinculado, principalmente, aos dados governamentais. Esse movimento demonstra como a tendência de valorização dos dados pode contribuir para a transparência

1 Doutora pelo Programa de Pós-Graduação em Educação (PPGEd-Unesp). Docente do Departamento (DCI) e do Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Estadual Paulista (Unesp). E-mail: silvana.vidotti@unesp.br

2 Doutoranda pelo Programa de Pós-Graduação em Ciência da Informação (PPGCI-Unesp). Bibliotecária da Universidade Tecnológica Federal do Paraná (UTFPR). E-mail: emanuelle@utfpr.edu.br

3 Doutor pelo Programa de Pós-Graduação em Ciência da Informação (PPGCI-Unesp). Docente do Centro Universitário Eurípedes de Marília (UNIVEM). E-mail: caio.coneglian@gmail.com

do poder público, bem como para aprimorar a eficiência dos serviços prestados à população. O movimento de dados abertos é uma tendência em todo o planeta e, quando alinhados aos interesses das pessoas, é capaz de melhorar o bem-estar de uma comunidade.

No âmbito científico, uma tendência cada vez mais valorizada é o movimento de acesso aberto, que visa dar transparência e abertura para os resultados científicos alcançados no processo da pesquisa. Tal movimento almeja que os pesquisadores demonstrem claramente os seus resultados de pesquisa, inclusive com a disponibilização livre dos dados coletados e resultados alcançados.

De forma ainda mais ampliada, a ciência aberta é outro movimento fortemente vinculado à disponibilização dos dados de pesquisa que, quando analisado sob a perspectiva da publicação, demonstra o potencial da disponibilização e do compartilhamento de dados de pesquisa. Com o apoio de repositórios de dados, periódicos de dados e artigos de dados, tal movimento tem conduzido a uma transformação no modo como as pesquisas são realizadas, com impacto nas diversas fases do processo científico, inclusive as agências de fomento que passaram a exigir um plano de gestão de dados de pesquisa.

A partir das tendências apresentadas, identifica-se uma valorização dos aspectos que envolvem dados em diferentes áreas. Desde dados governamentais até dados de pesquisa, passou a ser essencial a realização da gestão dos dados. Nesse sentido, é fundamental que todos os aspectos envolvendo a capacidade das máquinas localizarem, acessarem e interoperarem os dados sejam compreendidos e bem definidos.

Adicionalmente, há outros aspectos que devem ser observados em todo o processo da pesquisa quando se trabalha com dados de pessoas e comunidades específicas, considerando o ciclo de vida dos dados. Essas comunidades e os dados de pesquisas gerados que as envolvem necessitam ter um tratamento especial, que assegure a soberania sobre os dados, de modo a potencializar os benefícios e mitigar os impactos, bem como possibilitar tratamento adequado para a disponibilização, acesso e (re)uso. Portanto, este capítulo de livro tem como objetivo discutir os princípios FAIR e CARE para gestão de dados de pesquisas com seres humanos de comunidades específicas.

2. Princípios FAIR

Os princípios FAIR (*Findable*/Localizável, *Accessible*/Acessível, *Interoperable*/Interoperável e *Reusable*/Reutilizável) foram propostos em 2016, como um meio de tratar os aspectos computacionais envolvidos na disponibilização dos dados em diferentes contextos, dentre eles os dados de pesquisa. Tais princípios foram desenvolvidos compreendendo um cenário em que, na mesma intensidade que os

dados passam a ser importantes para o processo científico e para toda a sociedade, houve a dificuldade para que pessoas e aplicações computacionais fossem capazes de localizar e utilizar os dados disponíveis. Assim, "[...] os Princípios FAIR enfatizam especificamente o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente, além de apoiar sua reutilização por indivíduos." (WILKINSON *et al.*, 2016).

Na proposta inicial do FAIR, Wilkinson *et al.* (2016) apontam a importância das aplicações computacionais nos ambientes que armazenam os dados e destacam a necessidade de essas aplicações serem capazes de terem informações sobre as bases de dados, de forma interoperável e com acesso facilitado. Dessa forma, as aplicações computacionais podem auxiliar as pessoas a localizarem e utilizarem os dados disponíveis. Os autores relatam ainda que, em ambientes com um número muito grande de bases de dados, as pessoas dependem de aplicações computacionais para se relacionarem com tais dados.

Partindo dessa compreensão, e tendo como referência os princípios FAIR, baseados em *Australian National Data Service* (2020), Torino, Coneglian e Vidotti (2020) apresentam os princípios FAIR na Figura 1.

Figura 1: Princípios FAIR

Fonte: Torino, Coneglian e Vidotti (2020, p. 15).

A Figura 1 destaca características vinculadas a cada um dos princípios FAIR, demonstrando elementos que devem ser considerados ao realizar a disponibilização ou a publicação de dados.

De acordo com GOFAIR (2020), o princípio *Findable*/Localizável (**F**) pressupõe que para que um dado seja utilizado e/ou reutilizado é necessário que seja localizado, acionado, legível e processável por humanos e aplicações computacionais. Além disso, é necessário adotar identificadores persistentes para os dados, descrevê-los exaustivamente por meio de metadados enriquecidos e disponibilizá-los em infraestrutura indexada. O princípio *Accessible*/Acessível (**A**) reflete a capacidade de um conjunto de dados ser acessado e as especificações para fazê-lo, incluindo a utilização de protocolos de comunicação, autenticação, níveis de acesso e a persistência dos metadados, ainda que os dados não estejam mais disponíveis. O princípio *Interoperable*/Interoperável (**I**) visa otimizar a comunicação entre diferentes sistemas e a integração de diferentes conjuntos de dados. Para tanto, os dados e os metadados precisam ser legíveis e adequados a padrões e vocabulários reconhecidos, potencializando a ligação com outros padrões e incluir referências qualificadas. Por fim, o princípio *Reusable*/Reutilizável (**R**) visa otimizar o processo de reutilização dos dados. A reutilização trata de quão bem estão descritos os dados e os metadados, incluindo informações sobre os direitos de uso, a proveniência e o contexto dos dados, de modo a permitir que os dados sejam combinados e reutilizados por outras instâncias.

De acordo com FORCE21 (2020, tradução nossa):

> [...] por meio da definição e amplo apoio de um conjunto mínimo de princípios e práticas orientadores acordados pela comunidade, provedores de dados e consumidores de dados - tanto máquinas quanto humanos - poderiam descobrir, acessar, interoperar e reutilizar sensatamente, com a devida citação, as vastas quantidades de informações geradas pela ciência contemporânea com uso intensivo de dados.

Destaca-se que os princípios FAIR se aplicam ao tratamento dos dados, dos metadados e das infraestruturas visando maximizar a localização, o acesso, a interoperabilidade e o reuso de dados.

3. Princípios CARE

A infraestrutura tecnológica e a conectividade aumentam o valor dos dados, com isso, é imprescindível que haja princípios norteadores para os processos de coleta, armazenamento e disponibilização. Nesse sentido e considerando a soberania dos dados indígenas - apresentada por Stone e Calderon (2019, tradução nossa) como "[...] o direito dos povos indígenas e das nações de governar a coleta, propriedade e aplicação de seus próprios dados [...]" -, sobre os quais apenas os povos in-

dígenas possuem primazia para tomada de decisões, de acordo com seus interesses e valores, o *Global Indigenous Data Alliance* estabeleceu em 2018, na *International Data Week and Research Data Alliance Plenary*, os Princípios CARE para a Governança de Dados Indígenas.

CARE é um acrônimo em inglês para *Collective Benefit*, *Authority to Control*, *Responsibility*, *Ethics*; cuja tradução para o português é Benefício Coletivo, Autoridade para Controlar, Responsabilidade, Ética.

> [...] os 'Princípios CARE para Governança de Dados Indígenas' foram desenvolvidos pelo International Indigenous Data Sovereignty Interest Group da Research Data Alliance (RDA). Eles visavam capacitar os povos indígenas, mudando o foco da governança de dados de consulta para relações baseadas em valores que promovem a participação indígena equitativa em processos de reutilização de dados, o que resultará em resultados mais equitativos, bem como preservando relacionamentos construídos na confiança e respeito. (CARROLL *et al.*, 2020b, tradução nossa).

Tal iniciativa alicerça-se, de acordo com Global Indigenous Data Alliance (2019), na Declaração das Nações Unidas sobre os Direitos dos Povos Indígenas (NAÇÕES UNIDAS, 2008), que reconhece os direitos dos indígenas de autogovernança e autoridade sobre o seu patrimônio cultural. A "língua, conhecimento, costumes, tecnologias, recursos naturais e territórios" são considerados por eles como dados indígenas, muitas vezes expressos de forma oral e tidos como essenciais para o seu desenvolvimento e os seus direitos.

Dessa forma, tornou-se mister para os povos indígenas estabelecer princípios que possibilitem a tomada de decisões acerca de seus dados de forma indiscriminada, tendo em vista que o movimento mundial em torno dos dados abertos, quer sejam governamentais ou de pesquisa, não se compromete com os interesses supracitados e que o intercâmbio de dados favorecido pelo movimento e estruturado em princípios como o FAIR não mencionam características éticas, culturais e/ou de contextos históricos. De acordo com Carroll *et al.* (2020a), isso se faz diante da tensão da comunidade indígena entre proteger seus dados e interesses e apoiar iniciativas como a abertura, o compartilhamento de dados e a aprendizagem de máquina, objetivando que pesquisadores, gestores e usuários de dados sejam "justos e cuidadosos".

Portanto, os princípios CARE buscam estabelecer governança sobre os dados orientada pelas pessoas. "Esses princípios complementam os princípios FAIR existentes, encorajando movimentos abertos e outros movimentos de dados para con-

siderar as pessoas e o propósito em sua defesa e atividades." (GLOBAL INDIGENOUS DATA ALLIANCE, 2019, tradução nossa). Carroll *et al.* (2020a) enfatizam que os princípios CARE foram projetados para complementar os princípios FAIR, visando a inclusão dos povos indígenas para que possam ser implementados em conjunto; além disso, destacam que a governança de dados indígenas abrange a administração e o controle sobre os dados, o que contém os processos de coleta, armazenamento, análise, uso e reuso.

Os quatro princípios CARE são estruturados em 12 subprincípios a eles associados, conforme a Figura 2.

Figura 2: Os Princípios CARE para Governança de Dados Indígenas

Fonte: Carroll et al. (2020b, tradução nossa).

O primeiro princípio, *Collective Benefit*/Benefício Coletivo (**C**), estabelece que "Os ecossistemas de dados devem ser projetados e funcionar de forma a permitir que os povos indígenas se beneficiem dos dados." (GLOBAL INDIGENOUS DATA ALLIANCE, 2019, tradução nossa). Para tanto, **C1** - *For inclusive development and innovation*/Para desenvolvimento inclusivo e inovação - os governos e instituições devem apoiar o (re)uso de dados pelos povos indígenas e comunidades, com vistas à inovação, geração de valor e desenvolvimento local; **C2** - *For improved governance and citizen engagement*/Para melhorar a governança e o envolvimento do cidadão - os dados possibilitam o envolvimento entre governos, instituições e cidadãos,

proporcionam transparência e auxiliam no planejamento, avaliação e tomada de decisões, além de oferecer informações de interesse dos povos indígenas; **C3** - *For equitable outcomes*/Para resultados equitativos - os dados indígenas estão relacionados aos seus valores e podem se estender à sociedade, de forma que todos os dados gerados devem beneficiar os povos indígenas de forma justa.

O segundo princípio Authority *to Control*/Autoridade para Controlar (**A**) consiste em:

> Os direitos e interesses dos povos indígenas nos dados indígenas devem ser reconhecidos e sua autoridade para controlar esses dados deve ser autorizada. A governança de dados indígenas permite que os povos indígenas e órgãos governamentais determinem como os povos indígenas, bem como terras indígenas, territórios, recursos, conhecimentos e indicadores geográficos, são representados e identificados nos dados. (GLOBAL INDIGENOUS DATA ALLIANCE, 2019, tradução nossa).

No **A1** - *Recognizing rights and interests*/Reconhecimento de direitos e interesses - os povos indígenas devem ter reconhecidos os direitos e interesses nos seus dados e conhecimentos, o que se faz por meio do consentimento livre, prévio e manifesto na coleta, incluindo os usos, as políticas de dados e os protocolos utilizados na coleta; **A2** - *Data for governance*/Dados para governança - os povos indígenas devem exercer governança sobre seus dados, que devem estar a eles disponíveis e acessíveis; **A3** - *Governance of data*/Governança de dados - os povos indígenas podem desenvolver protocolos de governança e acesso aos seus dados, sobretudo aqueles concernentes ao conhecimento indígena.

O terceiro princípio, *Responsability*/Responsabilidade (**R**), estabelece que "Aqueles que trabalham com dados indígenas têm a responsabilidade de compartilhar como esses dados são usados para apoiar a autodeterminação e benefício coletivo. A responsabilidade requer evidências significativas e abertamente disponíveis desses esforços e benefícios para os Povos Indígenas" (GLOBAL INDIGENOUS DATA ALLIANCE, 2019, tradução nossa). Nesse sentido, **R1** - *For positive relationships*/Para relacionamentos positivos - o uso dos dados indígenas é possível quando pautado em relacionamentos de respeito, reciprocidade e confiança, de forma que o pesquisador seja o responsável por assegurar que os dados coletados, suas interpretações e uso garantam e respeitem a dignidade dos povos indígenas; **R2** - *For expanding capability and capacity*/Para expandir a competência e a capacidade - o uso de dados indígenas requer responsabilidade recíproca, na competência em dados junto aos povos indígenas e o desenvolvimento de infraestruturas digitais que possibi-

litem a coleta, a gestão, a segurança e a governança de dados; **R3** - *For indigenous languages and worldviews*/Para línguas e visões de mundo indígenas - tais recursos devem gerar dados baseados nas linguagens, experiências, valores, princípios e visões de mundo dos povos indígenas.

Como quarto princípio, *Ethics*/Ética (**E**), "Os direitos e o bem-estar dos povos indígenas devem ser a principal preocupação em todos os estágios do ciclo de vida dos dados e em todo o ecossistema de dados." (GLOBAL INDIGENOUS DATA ALLIANCE, 2019, tradução nossa). Que consiste em: **E1** - *For minimizing harm and maximizing benefit*/Para minimizar danos e maximizar benefícios - minimizar danos que podem ser oriundos de estigmas ou *déficits* relacionados aos povos indígenas, pautando a coleta, o tratamento e o uso em preceitos éticos, alinhados às estruturas éticas indígenas e com os direitos estabelecidos por meio da Declaração das Nações Unidas sobre os Direitos dos Povos Indígenas; **E2** - *For justice*/Por justiça - utilizar processos éticos de forma a tratar os desequilíbrios de poder e recursos, bem como a forma com que afetam os direitos indígenas e humanos; **E3** - *For future use*/Para uso futuro - a governança de dados deve considerar o potencial uso futuro pautado em bases éticas, valores e princípios dos povos indígenas e, ainda, expressar nos metadados a proveniência, o propósito, os direitos de uso, incluindo as limitações, as obrigações no uso e o consentimento.

4. #SejaJUSTOeCUIDADOSO: discussões e apontamentos

O *Global Indigenous Data Alliance* (GIDA)[4] expressa, por meio de "*#BeFAIRandCARE*", que os princípios FAIR e CARE são complementares, ao considerar tecnologias, propósitos e pessoas nos movimentos de dados abertos. Assim, enquanto o FAIR enfatiza, principalmente, os aspectos computacionais, dada a sua relevância para que as aplicações computacionais possam auxiliar os humanos diante do expressivo volume e complexidade dos dados; o CARE, enquanto princípios para a governança de dados indígenas, enfatiza, sobretudo, as pessoas e o propósito, considerando a relevância dos dados para o avanço, a autodeterminação e a soberania dos povos indígenas.

No que tange ao CARE, embora tenha se constituído para a governança de dados indígenas, Carroll *et al.* (2020a) destacam que indígenas, nações, pessoas e comunidades são atores nas sociedades globais contemporâneas. Portanto, os princípios CARE abordam aspectos relevantes para diferentes populações, como minorias sociais, comunidades e coletivos, que desejam ou necessitam manter diferentes níveis de tratamento e responsabilidade para o uso de seus dados. E, dentre esses

4 Disponível em: https://www.gida-global.org/care.

aspectos, os autores destacam: a privacidade, o uso, o reuso e a gestão, que podem se constituir em elementos para o estabelecimento de padrões, políticas e agendas.

Ao se considerar a atividade de pesquisa científica, é evidente a convergência de diferentes movimentos que determinam padrões, princípios e práticas gerais e de seus domínios, cujas peculiaridades devem ser atendidas. Stone e Calderon (2019) reafirmam que "Os princípios CARE certamente nos levam a considerar as pessoas refletidas nos dados e como nossas ações com eles podem impactá-las".

A necessidade de gerir dados de pesquisa desde o planejamento por meio do plano de gestão de dados de pesquisa, consiste em uma ação cada vez mais requerida aos pesquisadores por instituições de ensino e pesquisa, agências de fomento, repositórios e periódicos científicos e de dados. E, embora possa parecer uma burocratização do fazer científico, possibilita ao pesquisador e aos demais envolvidos em um projeto de pesquisa planejar e registrar os processos e decisões tomadas ao longo da investigação de modo a documentar cada etapa, o que possibilita uma gestão otimizada, considerando o ciclo de vida da pesquisa e dos dados de pesquisa.

Na maioria dos casos, as instituições elencam um conjunto de ferramentas e *templates* de apoio ao pesquisador na elaboração do plano de gestão de dados de pesquisa, que são elementos importantes à gestão, à disponibilização e ao uso futuro dos dados. Contudo, tais *templates* são genéricos e abertos, quando poderiam guiar o pesquisador no planejamento e execução das etapas da pesquisa já baseados em princípios norteadores, dentre os quais destacamos o FAIR e o CARE. Dessa forma, os elementos necessários são tratados durante o processo de pesquisa e asseguram que a abertura dos dados seja realizada de forma adequada.

Em todo o processo de gestão de dados de pesquisa, desde o planejamento, a coleta, a disponibilização e o reuso, é imprescindível atentar-se para a relevância do tratamento destes dados para que eles possam respeitar princípios tecnológicos que os tornem facilmente processáveis por aplicações computacionais, e, com isso, sejam reutilizados em infraestruturas robustas e por humanos. Por outro lado, precisam ser respeitados os princípios humanos vinculados às pessoas, propósitos e as consequências que a coleta, armazenamento e disponibilização dos dados de comunidades específicas podem trazer para o desenvolvimento das próprias comunidades.

Enfatizamos, a partir de Carroll *et al.* (2020a), que embora os princípios CARE tenham sido definidos para a governança de dados indígenas, cujo contexto é reconhecidamente relevante, é possível expandi-los para outras comunidades específicas que igualmente podem necessitar deles para que tenham governança e possam se desenvolver a partir dos seus dados. Nesse sentido, destacamos outras minorias sociais, como: quilombolas, comunidades ribeirinhas, de assentamentos sociais, de periferias e LGBTQ+.

Evoca-se aos pesquisadores, às instituições, às agências de fomento, aos governos e aos formuladores de políticas que o planejamento e a execução de pesquisas, bem como a abertura dos seus resultados sejam adequadamente realizados, sobretudo quando materializados em dados de pesquisa abertos. Além de pautar-se nos diferentes princípios estabelecidos e validados, dentre os quais destacamos o FAIR e o CARE.

Avanços no sentido de operacionalizar tais princípios de forma computacional são objeto de estudo, envolvendo os domínios, as comunidades e as instâncias correlacionadas. De acordo com Carroll *et al.* (2020b), o *Research Data Alliance*, por meio do *International Indigenous Data Sovereignty Interest Group* e do FAIR *Data Maturity Model Work Group,* já iniciou as discussões necessárias à operacionalização dos princípios CARE em conjunto com os princípios FAIR. Nessa perspectiva, destacam que um dos desafios consiste na necessidade de aplicação do CARE em todas as etapas do ciclo de vida dos dados. Dessa forma, podem ser tratadas adequadamente questões relacionadas à otimização de processos computacionais para localização, acesso e (re)uso dos dados, bem como a soberania e o tratamento equitativo dos dados.

Contudo, para que isso ocorra, além do tratamento adequado dos dados de pesquisa, é necessário que o pesquisador adote práticas cotidianas em todo o processo de pesquisa e em todo o ciclo de vida dos dados. Essa postura fará com que os processos ocorram de forma justa e cuidadosa.

Para tanto, no contexto da pesquisa científica e dos dados de pesquisa, deve-se adotar as práticas #SejaJUSTOeCUIDADOSO para ser:

a) JUSTO: disponibilizar dados de forma aberta e consoante aos princípios FAIR que favoreçam a localização, o acesso, a interoperabilidade e o uso, com:
   - utilização adequada e exaustiva de metadados nos conjuntos de dados e para a representação;
   - adoção de vocabulários padronizados;
   - adoção de identificadores persistentes;
   - escolha de ambiente digital adequado para a disponibilização e/ou publicação dos dados;
   - uso de formatos, protocolos e padrões abertos;
   - determinação de licença de uso;
   - manutenção de períodos de embargo razoáveis;
   - indicação adequada da proveniência dos dados;
   - versionamento dos dados;
   - preservação dos dados;

- reconhecimento de créditos, quando utilizados dados de pesquisa coletados por terceiros;
- utilização dos dados de acordo com o estabelecido na licença dos dados.

b) CUIDADOSO: respeitar os princípios CARE, considerando a hegemonia das comunidades específicas, suas visões de mundo, soberania e governança sobre os seus dados, com:
   - foco no desenvolvimento inclusivo;
   - estabelecimento de relações equitativas, de confiança, reciprocidade e respeito;
   - atendimento aos preceitos éticos e legais na coleta, tratamento, armazenamento e disponibilização dos dados;
   - identificação dos dados;
   - proteção aos direitos, interesses, valores e cultura;
   - participação na governança e controle sobre os seus dados;
   - melhoraria na representação dos dados;
   - capacitação para uso dos dados.

Ser justo consiste em formalizar os dados e disponibilizá-los seguindo boas práticas e princípios, para que possam ser de fato reutilizados por humanos e aplicações computacionais. E ser cuidadoso inclui e expande o aspecto anterior ao tratar adequadamente as pesquisas que envolvem seres humanos, sobretudo comunidades específicas e minorias sociais. Assim, as práticas de #SejaJUSTOeCUIDADOSO claramente devem estar presentes no fazer diário daqueles que atuam, direta ou indiretamente, com processos relacionados à coleta, à análise, ao tratamento, ao armazenamento e à disponibilização de dados, sobretudo, quando envolvem seres humanos.

5. Considerações finais

Os princípios FAIR e CARE buscam construir práticas de dados confiáveis, justas e responsáveis, tanto nos processos de gestão e de governança, como nos resultados e na qualidade dos conjuntos de dados que são disponibilizados.

Vale destacar que os princípios CARE estão envolvidos em todo o ciclo de vida dos dados, iniciando no plano de gestão de dados, perpassando os processos de coleta, representação, armazenamento e eventual disponibilização e reuso dos dados, respeitando os benefícios coletivos, a autoridade para controlar, a responsabilidade e a ética. Já os princípios FAIR, de modo análogo, também estão vinculados ao ciclo

de vida, porém enfocam a infraestrutura tecnológica para que os dados possam ser localizáveis, acessíveis, interoperáveis e reutilizáveis.

Dessa forma, recomenda-se a adequação dos planos de gestão de dados aos princípios FAIR e CARE, e a adoção pelos pesquisadores das práticas #SejaJUSTOeCUIDADOSO em todo o processo de pesquisa e no ciclo de vida de dados, especialmente em pesquisas relacionadas às pessoas de comunidades específicas para o tratamento equitativo dos dados.

Neste capítulo, inicia-se a discussão ao apontar caminhos que podem ser trilhados no trabalho com dados de pesquisa que envolvam seres humanos. Tais discussões podem ser aprofundadas, por exemplo, com base nas recomendações da Organização das Nações Unidas (ONU) para os direitos humanos, nos Objetivos do Desenvolvimento Sustentável (ODS) e nas legislações nacionais e internacionais para a gestão de dados pessoais.

7. Referências

AUSTRALIAN NATIONAL DATA SERVICE. FAIR **data training**. Disponível em: https://www.ands.org.au/working-with-data/fairdata/training. Acesso em: 09 jul. 2020.

CARROLL, S. R.; HOLBROOK, J.; LOVETT, R.; MATERECHERA, S.; PARSONS, M.; RASEROKA, K.; RODRIGUEZ-LONEBEAR, D.; ROWE, R.; SARA, R.; WALKER, J. D.; ANDERSON, J.; HUDSON, M. The CARE Principles for Indigenous Data Governance. **Data Science Journal**, v. 19, n. 1, p. 43, nov. 2020a. DOI: http://doi.org/10.5334/dsj-2020-043. Disponível em: https://datascience.codata.org/articles/10.5334/dsj-2020-043/#:~:text=The%20CARE%20Principles%20are%20a,value%20of%20data%20for%20reuse. Acesso em: 05 fev. 2021.

CARROLL, S. R.; HUDSON, M.; HOLBROOK, J.; MATERECHERA, S.; ANDERSON, J. **Working with the CARE principles: operationalising Indigenous data governance**. 2020b. Disponível em: https://www.adalovelaceinstitute.org/blog/care-principles-operationalising-indigenous-data-governance/. Acesso em: 06 fev. 2021.

FORCE21. **Guiding principles for findable, accessible, interoperable and reusable data publishing version b**1.0. 2020. Disponível em: https://www.force11.org/fairprinciples#Annex1-1. Acesso em: 07 fev. 2021.

GLOBAL INDIGENOUS DATA ALLIANCE. CARE **Principles for Indigenous Data Governance**. 2019. Disponível em: https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf. Acesso em: 05 fev. 2021.

GOFAIR. FAIR **principles**. 2020. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 07 fev. 2021.

NAÇÕES UNIDAS. **Declaração das Nações Unidas sobre os direitos dos povos indígenas**. Rio de Janeiro, 2008. Disponível em: https://www.un.org/esa/socdev/unpfii/documents/DRIPS_pt.pdf. Acesso em: 05 fev. 2021.

STONE, P.; CALDERON, A. **[Spotlight]** CARE **Principles**: unpacking indigenous data governance. 2019. Disponível em: https://opendatacharter.medium.com/spotlight-care-principles-f475ec2bf6ec. Acesso em: 06 fev. 2021.

TORINO, E.; CONEGLIAN, C. S.; VIDOTTI, S. A. B. G. Estruturas de representação para reuso de dados no contexto da ecologia de pesquisa: CRIS institucional. **Informação & Informação**, Londrina, v. 25, n. 3, p. 1-27, out. 2020. DOI: http://dx.doi.org/10.5433/1981-8920.2020v25n3p1. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/41946. Acesso em: 06 fev. 2021.

WILKINSON, M., DUMONTIER, M., AALBERSBERG, I. et al. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, 160018, mar. 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 06 fev. 2021.

Como citar com o DOI individual

VIDOTTI, Silvana Aparecida Borsetti Gregorio; TORINO, Emanuelle; CONEGLIAN, Caio Saraiva. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 201 - 214. DOI: 10.22477/9786589167242.cap15

Um modelo de implementação para a internet de dados & serviços FAIR

Luís Fernando Sayão[1] e Luana Farias Sales[2]

1. Introdução

Não há nenhuma novidade em afirmar que a vasta e crescente quantidade de dados e informação que se propaga por toda a sociedade contemporânea vai remodelando profundamente o seu *modus vivendi* em todas as suas dimensões, ativadas por sistemas tecno-sociais: lazer, educação, administração pública, negócios, cuidado com saúde, expressão cultural e, sobretudo, as interlocuções pessoais. Essa "infoesfera" intensivamente conectada e planetariamente distribuída se torna possível por meio do avanço vertiginoso das tecnologias de computadores e de redes – e de sua materialidade digital - que proporcionam a criação, captura, cópia, transmissão, compartilhamento e armazenamento massivo de informação de forma fácil e um custo muito baixo. (national research council, 2015). É de se esperar que essas transformações totalizantes se sobreponham de forma contundente aos processos de construção do saber científico.

Não importando o ponto de observação, o que se constata é que a pesquisa científica – por essa tendência global aliada às suas conexões imanentes com os sistemas técnicos - está produzindo um enorme e crescente fluxo de dados digitais. Inúmeros sensores instalados nos mais diversos dispositivos que vão de longínquos satélites, aceleradores de partículas, sequenciadores automáticos de dna, até em despretensiosos implantes médicos, permitem que dados sejam capturados em uma quantidade sem precedentes em todos os domínios científicos, das ciências exatas às humanidades, arte e cultura.

Em face dessas constatações, a gestão de dados de pesquisa se configura, atual-

1 Doutor em Ciência da Informação pelo ppgci ibict-ufrj, pesquisador da cnen, docente do ppgci ibict-ufrj, luis.sayao@cnen.gov.br

2 Doutora em Ciência da Informação ppgci ibict-UFRJ, docente do ppgci ibict-ufrj, luanasales@ibict.br

mente, como um foco de interesse e um dos maiores desafios para as organizações de pesquisa. Como desdobramento, a gestão e curadoria de dados, em escala planetária, se destacam com proeminência no cenário da pesquisa do século XXI, bem como a ubiquidade das tecnologias digitais para a coleta, análise e arquivamento de dados em quase todos os domínios disciplinares (MAYERMIK, 2012). Assim sendo, as instituições de pesquisa, em gradações distintas, estão reconceituando a gestão de dados e a identificando como parte integrante dos processos de pesquisa, reconsiderando ou ampliando as suas estratégias de tratamento dos dados, implementando plataformas de gestão e curadoria, adquirindo ferramentas de análises e desenvolvendo programas de capacitação para as suas equipes.

São muitas as motivações para a implementação de novas modalidades de serviços de informação que apoiam a gestão de dados nos ambientes acadêmicos e de pesquisa, dentre elas está a necessidade de apoiar as atividades de pesquisa, acelerar o progresso científico e a inovação por meio do intenso compartilhamento e colaboração em âmbito local e internacional (MUSHI, 2020). Porém, podemos identificar o reuso e a reprodutibilidade como principais objetivos da gestão de dados e importantes parâmetros de avaliação, a partir dos quais outros benefícios são constituídos. Existem muitas motivações para armazenar e preservar dados, mas a razão primordial é o reuso, enfatiza Borgman (2007) e a reprodutibilidade.

O planejamento, desenvolvimento e implantação de plataformas de gestão de dados de pesquisa, devido ao número de variáveis que precisam ser equacionadas, são problemas complexos e multifacetados. Precisam ser articulados em torno de fluxos de trabalho, de domínios disciplinares específicos, parâmetros informacionais, tecnológicos, políticos, éticos e legais, de sustentabilidade e de expertise numa odisseia marcada por constantes mudanças, cujo signo é a heterogeneidade.

Este ambiente complexo pode ser um terreno adequado para a adoção dos Princípios FAIR como horizonte para a implementação de serviços de gestão que tornem os dados de pesquisa encontráveis, acessíveis, interoperáveis, para que possam ser reusados por longo prazo, criando-se, dessa forma, condições para a transição de uma pesquisa autocontida, para uma pesquisa mais aberta, em rede e cooperativa, que, ao mesmo tempo, atenda requisitos disciplinares que beneficiem comunidades de culturas e restrições específicas. "Os Princípios FAIR não são mágicos e não representam uma panaceia, mas eles orientam o desenvolvimento de infraestruturas e ferramentas que tornam todos os objetos de pesquisa reusáveis de forma otimizada igualmente para máquinas e pessoas", enfatizam Barend Mons e seus colaboradores (2017, p.55), fundadores desse movimento. Entretanto, o alinhamento e implementação dos princípios FAIR em uma instituição de pesquisa

exigem investimentos financeiros, mudanças culturais, treinamento e a construção de uma infraestrutura técnica (GRAAF; WAAIJERS, 2011), fatores que podem ser aglutinados em torno do conceito de "plataforma de gestão de dado de pesquisa". Esse tipo de plataforma tem o potencial de operacionalizar as diversas camadas de gestão e estabelecer uma crescente infraestrutura de serviços informacionais, científicos e computacionais na direção de aplicar os Princípios FAIR em objetos de pesquisa, chamado processo de FAIRificação, sejam eles dados propriamente ditos ou algoritmos, códigos, procedimentos, fluxos de trabalho (*workflows*) ou outros dispositivos físicos ou conceituais que levam aos dados.

Tentando equacionar essa diversidade, o presente trabalho tem como objetivo apresentar uma arquitetura genérica para apoiar o projeto de plataformas de serviços de dados, definindo, realinhando, agregando e articulando os vários módulos conceituais – diretrizes, políticas, serviços, ferramentas, infraestruturas, dentre outros- em torno de um modelo de camadas que, como blocos de construção, podem ser ajustados de acordo com a profundidade, alcance e filosofia de cada instituição ou disciplina. O modelo pretende se constituir numa possível escala para mensuração do nível de maturidade dos projetos de serviços de gestão. A arquitetura proposta tem como horizonte tornar os dados aderentes aos Princípios FAIR, abrindo a perspectiva para que um número crescente de aplicações e serviços possam linkar e processar dados FAIR, realizando a ideia de "Internet de Dados & Serviços FAIR" – IFDS na sigla em inglês – que se desdobram em diversos benefícios para os vários *stakeholders* envolvidos.

Para delineamento dos elementos da arquitetura proposta, foi tomado como metodologia a análise da literatura da área, com ênfase especial nos artigos, relatórios, manuais e projetos de infraestrutura de dados elaborados por pesquisadores e instituições de pesquisa.

2. Algumas considerações sobre os princípios FAIR e sua implementação

A noção de gerenciamento apropriado de dados de pesquisa, idealizada de forma que seja capaz de maximizar as oportunidades de descoberta e o reuso eficientes de resultados de pesquisa por humanos e máquinas, não é propriamente uma novidade já que está presente há décadas nos domínios da pesquisa científica, em particular pelas comunidades de web semântica e engenharia de ontologias. Neste percurso, muitas opções de implementação já foram realizadas por comunidades pioneiras para associar o gerenciamento de dados com a noção de "acionabilidade da máquina". Os Princípios FAIR podem ser vistos como uma síntese desses esforços anteriores e surgiram da materialização de uma visão, de múltiplas partes interessadas, de uma infraestrutura consolidada de suporte ao reuso de dados que podem ser processados

por computadores (WILKINSON *et al,* 2016), que foi posteriormente cunhada de "Internet de Dados & Serviços FAIR" (JACOBSEN *et al,* 2020; MONS *et al*, 2017).

Os Princípios dos Dados FAIR preconizam que todos os produtos de pesquisa devam ser encontráveis, acessíveis, interoperáveis e assim reusáveis por seres humanos e por máquinas, expressando a expectativa dos pesquisadores em relação aos recursos de dados na ciência atual, e oferecendo um guia para produtores e publicadores de dados para que eles naveguem com mais segurança e objetividade em torno da complexidade inerente à gestão de dados de pesquisa. O foco primordial dos Princípios está em assegurar que os dados possam ser reusados, tanto por humanos quanto por máquinas (Inteligência Artificial), em pesquisas subsequentes e reinterpretados transversalmente acelerando a interdisciplinaridade e a inovação, tornando-se deste modo ainda mais valiosos; e ainda maximizando o valor agregado obtido pelo desenvolvimento das publicações acadêmicas que tenham como substrato as tecnologias digitais e de redes (WILKINSON *et al*, 2016). Nesta direção, os Princípios FAIR delineiam considerações que se inserem nos ambientes de publicação contemporânea de dados de pesquisa e estão relacionadas com o deposito, exploração, compartilhamento e reuso desses recursos por processos manuais e automatizados. Assim sendo, eles descrevem as características que os recursos de dados, ferramentas, vocabulários e infraestruturas devem ter para apoiar a descoberta e a reutilização por outros *stakeholders* em empreendimentos subsequentes, de agora e do futuro.

De forma diferente das iniciativas moldadas por domínios disciplinares que estabelecem práticas específicas para a gestão e arquivamento de dados, FAIR "descreve princípios de alto nível, concisos, independentes de domínio que podem ser aplicados a um amplo espectro de produtos de pesquisa" (WILKINSON *et al*, 2016, p.2), podendo ser, porém, "base para o desenvolvimento de padrões comunitários [e disciplinares] flexíveis" (BOECKHOUT; ZIELHUIS; BREDENOORD, 2018, p. 932). Mesmo diante dessa neutralidade, padrões conhecidos, como o Resource Description Framework (RDF) da W3C em conjunto com ontologias formais, são atualmente soluções frequentemente aplicadas para interoperabilidade e compartilhamento de informação e conhecimento que atendem aos requisitos FAIR, especialmente ao nível de metadados (MONS *et al*, 2017, p.51).

Na qualidade de uma concepção de alto nível, a adoção dos Princípios FAIR precede as escolhas de implementação, que não recomendam nenhuma tecnologia específica ou solução para tal, o que não se constitui uma norma, padrão ou especificação. Porém, oferecem um conjunto de orientações para a gestão voltada para o reuso de recursos digitais de pesquisa. Os elementos dos quatro princípios FAIR são relacionados, porém independentes e separáveis, podendo ser implementados em

qualquer combinação e de forma incremental na medida em que o ambiente de publicação evolua na direção de níveis maiores de "FAIRness". A importância e o grau de implementação de cada princípio podem depender de prioridades e maturidade de cada comunidade no uso de determinados objetos de pesquisa (HONG *et al*, 2020). Estas características contribuem para a ampla adoção dos princípios, posto que comunidades específicas, incluindo as não pertencentes ao mundo científico, podem implementar suas próprias soluções FAIR, permitindo que elas possam ser reconfiguradas ao longo do tempo para acompanhar a evolução das tecnologias subjacentes (JACOBSEN *et al,* 2020). Assim sendo, é preciso reconhecer que diferentes disciplinas requerem diferentes tipos de soluções técnicas para alcançar os mesmos benefícios dos dados FAIR.

É de importância crucial destacar que a aplicação dos princípios FAIR extrapola os dados de pesquisa no seu sentido mais convencional. No escopo mais restrito das práticas científicas e metodológicas, os princípios FAIR devem se estender também aos algoritmos, códigos, ferramentas, metodologias e *workflows,* objetos que conduzem a obtenção dos dados e que, se bem documentados, permitem o rastreamento da proveniência desses ativos. Assim, eles precisam ser identificados, descritos e reusados, como os dados.

Todos os objetos digitais de pesquisa – dos dados aos *pipelines* analíticos – se beneficiam da aplicação desses princípios, posto que todos os componentes dos processos de pesquisa devem estar disponíveis para garantir transparência, reprodutibilidade e reusabilidade" (WILKINSON *et al*, 2016, p.1).

Esta característica aproxima os princípios FAIR dos pressupostos da ciência aberta, cujas considerações precisam ir além das publicações convencionais.

A ideia primordial de implementação de uma Internet de Dados & Serviços FAIR não se realiza por si só. Para tal, é necessário um processo de gestão de dados que possa efetivamente ir agregando valor ao longo do tempo. O grau de aderência dos produtos de pesquisa aos princípios FAIR está vinculado ao alcance e profundidade da gestão a que eles estão submetidos. Isto pressupõe a necessidade de um arcabouço de várias camadas – científica, tecnológica, informacional e de governança, que enderecem os inúmeros problemas éticos, metodológicos e organizacionais que se interpõem entre os fluxos de compartilhamento, integridade, reprodutibilidade, prestação de contas da pesquisa, bem como as novas necessidades e oportunidades de análise e reanálise em larga escala (WILKINSON *et al*, 2016).

3. Princípios FAIR x gestão de serviços

A efetivação dos princípios FAIR se dá pelos graus variados de ações aplicadas aos dados pelo conjunto de serviços de gestão de dados disponibilizados princi-

palmente pelas diversas plataformas disciplinares. Esses conjuntos de serviços de FAIRificação são captados pelo modelo em três categorias: informacionais, computacionais e científicos. Boeckhout, Zielhuis e Bredenoord (2018) deixam essa relação bem clara na sua análise para a área de genoma que, entretanto, pode ser generalizada.

- O princípio da **Encontrabilidade** estipula que os dados têm que ser fáceis de se achar por seres humanos e máquinas. Dessa forma, os dados têm que ser **identificados, descritos e registrados ou indexados de uma maneira clara e inequívoca para que sejam localizados e seu conteúdo compreendido por exploradores humanos e computacionais**. Em termos de serviços, isto significa que numa coleção de dados deve ser assinalado um **identificador** único e persistente; que as principais características da coleção sejam sistematicamente especificadas, idealmente utilizando formatos padronizados; e que ela seja depositada ou indexada em um dispositivo público tal como um arquivo ou centro de dados ou um repositório disciplinar ou institucional, o que enfatiza a necessidade dos serviços informacionais e infraestruturas de gestão. Metadados significativos e acionáveis por máquina são essenciais para a encontrabilidade automática de conjuntos de dados e serviços relevantes e, portanto, são um componente essencial do processo de FAIRificação (JACOBSEN *et al*, 2020).
- O princípio da **acessibilidade** preconiza que os objetos de pesquisa sejam acessíveis preferencialmente por meios da implementação, quando apropriados, de protocolos automatizados de recuperação de dados; estipula também que os dados estejam disponíveis de acordo com procedimentos e condições claras e bem definidas. Estas condições envolvem o estabelecimento de processos de autenticação e autorização que estejam alinhados às políticas da organização e à cultura disciplinar, e ainda, às especificidades dos dados – por exemplo, nível de sensibilidade. Como mantra FAIR, que deve ser trabalhado pelos serviços de gestão de dados, temos que os metadados devem estar incondicionalmente acessíveis mesmo que os dados não estejam, ou deixaram de estar disponíveis.
- O princípio da **Interoperabilidade** é o mais difícil de se implementar (HONG *et al*, 2020), posto que está condicionado a um alto grau de padronização em todas as suas articulações. Em termos gerais, quando dois ou mais recursos digitais estão relacionados ao mesmo tópico ou entidade, deve ser possível para a máquina mesclar as informações de cada um dos recursos numa visão unificada e mais rica desse tópico ou entidade; da mesma forma, quando

uma entidade digital é capaz de ser processada por um serviço on-line, uma máquina deve ser capaz de detectar automaticamente essa conformidade e facilitar a interação entre os dados e essa ferramenta. Isso requer que o significado (semântica) de cada recurso participante - sejam eles dados e/ou serviços - seja claro (JACOBSEN *et al*, 2020). Neste sentido, o princípio da Interoperabilidade pressupõe que dados e metadados sejam conceitualizados, expressos e estruturados por meio de padrões de ampla aceitação, publicados, rastreáveis e acessíveis (ou seja, também FAIR). Para alcançar esse objetivo, a implementação desse princípio compreende uma rigorosa aplicação de padrões técnicos e semânticos em todos os níveis – científicos, computacionais e informacionais – como por exemplo em termos de variáveis, protocolos, formatos de arquivo, ontologias e fluxos de trabalho.

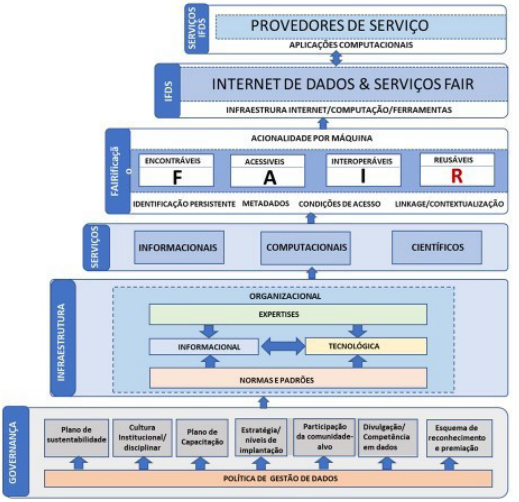
- O princípio da **Reusabilidade** é uma decorrência dos princípios anteriores e reforça os pontos importantes preconizados por eles como a descrição detalhada das características dos dados para seres humanos e computadores, incluindo a proveniência, de acordo com os padrões relevantes para as comunidades de domínios específicos. Isto permite que uma máquina possa decidir: se um recurso digital deve ser reusado – ou seja, se ele é relevante para a tarefa em questão; se um recurso digital pode ser reutilizado e em que condições – ou seja, se o recurso preenche as condições de reutilização; e a quem dar o crédito caso o recurso seja reusado. O grau de reusabilidade coloca em destaque a necessidade de licenças apropriadas as condições de reuso requerido.

4. Descrição do modelo proposto

A gestão de dados de pesquisa tem muitas faces, mas nenhuma delas consegue de forma isolada explicar completamente a complexidade intrínseca dos seus processos. Partindo desse ponto, o modelo procura representar os diferentes aspectos da gestão de dados de pesquisa sem perder de vista a natureza interrelacionada da dinâmica das atividades que se desenrolam num ambiente científico intensivo de dados cujo objetivo é tornar os dados FAIR. Como uma abstração conveniente da realidade que se quer compreender, um modelo é uma criação cultural, um "mentefato", destinada a representar uma realidade, ou alguns dos seus aspectos, a fim de torná-los descritíveis qualitativa e quantitativamente e, algumas vezes, observáveis (SAYÃO, 2001). A partir desse ponto, decidiu-se dividir o modelo em seis camadas representacionais: 1) governança, onde são discutidos os princípios norteadores do projeto de serviços de gestão de dados; 2) as infraestruturas técnicas onde se

incluem também as categorias de expertises necessárias;3)os serviços informacionais, computacionais e científicos; 4) os resultados da efetivação desses serviços manifestados pela FAIRificação dos dados; 5) que, por sua vez, consolida-se em um ambiente global e compartilhado, conceituado como Internet de Dados e Serviços FAIR (SALES *et al*, 2020), onde 6) Provedores de Serviços, por meio de aplicações computacionais, oferecem serviços de naturezas diversas. A figura 1 apresenta uma visão geral dos componentes agrupados em camadas e de suas interrelações, que a seguir são discutidos.

Figura 1 – Modelo de implementação para Internet de dados e serviços FAIR



Fonte: elaborada pelos autores

4.1 Governança de Dados de Pesquisa: planejamento, política, institucionalização e sustentabilidade

A configuração organizacional na qual a gestão de dados é realizada pode variar em relação à vários aspectos, como a intensidade de apoio à gestão e o nível de investimentos aplicados. Algumas instituições como centros referenciais de dados científicos e agências estatísticas governamentais podem estar inteiramente dedicadas à gestão de dados, tendo-a como finalidade principal; em outras configurações, a gestão de dados é parte de uma atividade mais ampla que se conecta à outras atividades de pesquisa, como, no caso das universidades (NATIONAL RESEARCH COUNCIL, 2015), cuja atividade de gestão de dados é decorrente de suas funções de ensino, pesquisa e extensão. Porém, mesmo no contexto acadêmico, são mui-

tas as formas de planejar e executar as tarefas de gestão de dados que variam de acordo com referências objetivas como graus de investimento, sistemas técnicos disponíveis, volume e tipo de dados e de como a gestão de dados está integrada aos seus fluxos de trabalhos e processos; e com percepções mais subjetivas como cultura disciplinar e prestígio acadêmico. No presente modelo, esses parâmetros são equacionados por um nível mais administrativo compreendido pelo termo "governança de dados". Num patamar mais conceitual, governança de dados delineia os princípios, políticas e estratégias que são comumente adotados num ambiente que necessita de um programa de gestão de dados coerente; delineia também as ações, funções e papéis que são necessários para implementar essas políticas e estratégias. No âmbito de uma instituição de pesquisa, os princípios, operacionalizados pela gestão governam todo o ciclo de vida dos dados – da conceitualização ao arquivamento e possível descarte. O processo de governança de dados trata os dados não somente em seu aspecto espacial, mas também ao longo da sua dimensão temporal (SOLOMONIDES, 2019), este requisito implica uma ampliação do grau de complexidade e envergadura dos comprometimentos da governança.

Este arcabouço estruturante é necessário posto que dados de pesquisa digitais só podem ser gerenciados e preservados adequadamente ao longo do tempo por meio de um compromisso institucional sustentado (MAYERMIK, 2012, p.1). Em certa medida, a consolidação dos serviços de gestão de dados reflete o nível de aceitação organizacional incorporada a eles e o grau de planejamento das várias ações necessárias: orçamento sustentável em vigor, política de dados apropriada, conexão orgânica com as comunidades-alvo, conformidade com os códigos éticos e legais, alinhamento com os objetivos estratégicos institucionais e uma estratégia de desenvolvimento que considere os percursos possíveis para cada instituição. É necessário considerar também a inevitabilidade do fato de que as infraestruturas tecnológicas para acessar, interpretar e preservar a informação digital está continuamente evoluindo; antecipar esses problemas e desenvolver estratégias para mitigá-los é uma atividade relevante para os compromissos de governança (NATIONAL RESEARCH COUNCIL, 2015). Sobre esses pilares podem ser desenvolvidos serviços avançados de dados que possam apoiar apropriadamente todo o ciclo de vida desses ativos informacionais de acordo com os interesses dos vários *stakeholds* envolvidos. Considerando essas questões, propomos os seguintes enfoques como parte do modelo:

- *Política de Gestão de Dados da Instituição* – Estabelece os fundamentos, diretrizes e compromissos da instituição concernentes à gestão, uso, propriedade, conformidade aos códigos éticos e legais, aderência às políticas das

agências de fomento, políticas nacionais de ciência, tecnologia e inovação, às orientações e práticas internacionais e, por fim, mas de importância crítica, à cultura, às práticas e às idiossincrasias das comunidades e domínio disciplinares: uma política de gestão de dados de pesquisa abrangente deve também identificar as responsabilidades de cada um dos atores – biblioteca, laboratórios, tecnologia da informação, administração etc. – posto que a gestão de dados envolve diferentes setores da instituição (MUSHI, 2020) e o projeto é considerado como parte das atividades de pesquisa da instituição. É necessário enfatizar que o processo de desenvolvimento de uma política institucional de gestão de dados requer uma consulta extensiva a todos os agentes envolvidos e a aprovação das comunidades e organizações científicas relevantes (WILSON *et al*, 2011). As orientações da política devem permear todo o ciclo da gestão. "Políticas podem ser um importante fator motivador para dados FAIR e outros objetos de pesquisa (*software*, *workflow*, modelos, protocolos etc.). Portanto é essencial que esforços "*botton-up*" baseados na comunidade sejam combinados com políticas com enfoque "*top-down*", completam Hong e seus colaboradores (2020).

- *Cultura Institucional/disciplinar* – A implantação de uma plataforma de serviços de gestão de dados de pesquisa deve ser precedida de uma análise de requisitos que considere o contexto e a cultura institucional, comunitária e disciplinar e suas características únicas. Espera-se que este processo ajude a definir uma carteira de serviços de gestão de dados mais efetiva para apoiar as práticas de pesquisa da instituição e de suas comunidades (MUSHI, 2020; COATES, 2014; REED, 2015). É importante também reconhecer que algumas disciplinas requerem diferentes tipos de soluções técnicas para obter os mesmos benefícios dos dados FAIR (HONG *et al*, 2020).
- *Plano de sustentabilidade* – Um dos grandes desafios de um programa de implementação de uma infraestrutura de gestão de dados é assegurar que cada fase do projeto seja sustentável como um serviço contínuo ao longo do tempo (WILSON *et al*, 2011). Uma vez que a gestão de dados de pesquisa é reconhecida como algo necessário para as atividades de pesquisa, os seus custos devem ser estimados e suas fontes de financiamento – especialmente as perenes – identificadas. Desta forma, um projeto de implementação de serviços de gestão de dados de pesquisa precisa estar associado a um plano de sustentabilidade que delineie um comprometimento possível com o agora e com o futuro. A criação e o comprometimento com uma estratégia de longo prazo para os serviços podem revelar com mais clareza os recursos necessários à continuidade dos serviços e das infraestruturas necessárias para tal.

Isso, portanto, pode incluir, um plano de sucessão (MUSHI, 2020).

- *Divulgação/Competência em dados* – Para a implementação de um ambiente de pesquisa FAIR é necessário que as comunidades envolvidas desenvolvam uma compreensão compartilhada do que está circunscrito pelo conceito FAIR e pelos Princípios. De uma forma geral, os pesquisadores e outros *stakeholders* têm baixo nível de percepção sobre a importância das práticas de gestão de dados e das exigências de gestão e compartilhamento das agências de fomento e dos compromissos de depósito dos dados firmados com os editores científicos, além das questões éticas e legais envolvidas na publicação dos dados. Por exemplo, em relação ao conceito FAIR, Hong *et al* (2020) observam que o pesquisador não sabe o que é dado FAIR e muitas vezes acha que é o mesmo que dado aberto. Isto indica que é necessário planejamento e ações de divulgação e de conscientização que tragam à tona essas questões. Um programa de divulgação nesta direção deve contemplar a elaboração de material didático (cartilhas e guias), cursos, eventos, oficinas entre outros.
- *Conhecimento/participação da comunidade-alvo* – Como criadores e usuários de dados de pesquisa, o engajamento dos pesquisadores é crucial no desenvolvimento de serviços de gestão de dados. O provisionamento de qualquer serviço precisa ser baseado numa compreensão próxima dos padrões e fluxos das pesquisas que se desenvolvem na instituição, das suas motivações, características e prioridades. Assim sendo, a definição precisa dos requisitos dos serviços necessita ser estabelecida com o comprometimento e a contribuição da comunidade de pesquisadores, sem essas considerações as características dos serviços podem não estar em harmonia com os objetivos dos pesquisadores. A comunidade deve ser acompanhada nas mudanças de interesse sobre os dados, e a participação dela no desenvolvimento e escolha de padrões compartilháveis para as práticas e para infraestruturas FAIR deve ser reconhecida e institucionalizada. A proximidade, interação e alinhamento das comunidades com as organizações nacionais e internacionais que lidam diretamente com a gestão de dados FAIR como GO FAIR, RDA, CODATA, DCC e outras, devem ser incentivados.
- *Plano de capacitação* – Para oferecer serviços completos em gestão de dados, as bibliotecas precisam ter pessoal tecnologicamente qualificado ou aumentar muito as oportunidades de treinamento tecnológico para o pessoal existente. (TENOPIR *et al.*, 2012). Sustentabilidade humana é crítica para assegurar a continuidade e consistência da oferta de serviços ao longo do tempo. Entretanto, poucos programas formais em estudos informacionais incluem em seus currículos gestão de dados, dessa forma, os gestores de dados de

pesquisa são normalmente treinados em serviço nas disciplinas específicas onde trabalham (BORGMAN, 2007, p.155) .

- *Estratégia/níveis de implantação* – O desenvolvimento e a implantação de infraestrutura de gestão de dados, além de muitos recursos, requerem tempo para alcançar sua plena maturidade e espelharem as demandas das comunidades científicas, isto implica na necessidade do estabelecimento de níveis de implantação de infraestruturas e serviços. As bibliotecas de pesquisa, por exemplo, têm, em muitos casos e de forma proativa, procurado suprir as necessidades de gestão de dados para suas comunidades de usuário. Frequentemente isso acontece sem aporte financeiro adicional destinado ao desenvolvimento e disponibilização de serviços de dados. Assim sendo, as bibliotecas têm que começar numa escala mais simples, construindo uma base sobre a qual pode desenvolver serviços mais sofisticados (ERWAY *et al*, 2016, p.5), começando com serviços básicos que exijam apenas recursos da própria biblioteca, até alcançarem serviços mais complexos que exijam alto nível de compromisso institucional e mais recursos financeiros, tecnológicos e humanos (KOUPER *et al*, 2017).
- *Recompensa e reconhecimento* – A gestão de dados de pesquisa consome tempo, recursos e exige grande dedicação do pesquisador, entretanto, esse esforço raramente é percebido pelo sistema de recompensa acadêmico, exceto quando linkado com publicações em periódicos científicos. Portanto, para incentivar essa nova tarefa dos pesquisadores e destacar sua importância é essencial que ela seja apropriadamente reconhecida e que seja considerada nos critérios de avaliação, promoção e de contratação.

4.2. Infraestruturas de Dados de Pesquisa

Infraestrutura é uma noção de grande amplitude e multidimensional. Ela pode ter uma conotação técnica, legal, organizacional e, em muitos casos, é imprescindível considerar também os aspectos sociais, culturais e políticos. De fato, é assim no domínio da ciência: o projeto de infraestrutura de pesquisa é simultaneamente uma questão tecnológica, uma questão de identificação das necessidades da pesquisa em áreas disciplinares específicas e uma questão política. Essa ótica mais geral se aplica às infraestruturas institucionais de gestão de dados de pesquisa que precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

Assim, da mesma forma que as instituições devem providenciar infraestruturas básicas para a pesquisa – tais como laboratórios, instrumentação, computação

de alto desempenho, redes, reagentes e muito mais – elas devem também tomar medidas para uma gestão adequada dos dados. Isto pressupõe um amplo espectro de atividades gerenciais, tecnológicas e informacionais que inclui profissionais de informação treinados para apoiar pesquisadores no planejamento e gestão de seus dados, no acesso a dispositivos de armazenamento seguro e *backups* durante o desenvolvimento do projeto e disponibilidade de plataformas de acesso e de preservação de longo prazo, necessárias após o fim da pesquisa (STRASSER, 2015); é imprescindível também um corpo de normas, padrões e boas práticas que permitam, principalmente, uma interlocução em níveis variados dos sistemas e serviços, tanto local quanto global, que pode ser traduzida por interoperabilidade.

Quando comparamos a publicação acadêmica tradicional com a publicação de dados verificamos que as infraestruturas subjacentes à publicação acadêmica criam uma ponte epistemológica entre disciplinas tendo como ponto agregador as bibliotecas de pesquisa que selecionam, coletam, organizam e tornam acessíveis publicações de todo tipo e de todas as áreas. Por sua natureza, as instituições sociais trabalham para estabilizar práticas particulares e formas de conhecimentos. Em certo sentido, as instituições são infraestruturas sociais em si mesmas. Nessa direção, as infraestruturas técnicas estão entrelaçadas com as infraestruturas sociais das instituições, muitas vezes mediadas por padrões, protocolos, documentos e artefatos que ligam os aspectos sociais e técnicos das infraestruturas (LEONARDI, 2010). Entretanto, não existe ainda infraestrutura dessa magnitude para os dados. Algumas poucas áreas têm mecanismos consolidados para publicar dados; outras estão nos estágios de desenvolvimento de padrões e práticas para agregar seus dados e torná-los mais amplamente acessíveis. "A falta de infraestrutura para dados amplifica a descontinuidade na publicação acadêmica", confirma (BORGMAN, 2007, p. 155).

Os arcabouços infraestruturais voltados para a gestão de dados são diversos e fragmentados em termos de fluxos, complexidade, aplicação e topologia, e organizados de forma diferente pelas várias disciplinas e em diferentes países (GRAAF; WAAIJERS, 2011). Contudo, crescentemente as infraestruturas moldam os padrões e as práticas da gestão de dados. Diante desse fato, o conhecimento sobre a origem, domínio disciplinar, grau de processamento, sistemas de coleta, *workflows* etc. parecem ser de importância crítica na concepção de infraestruturas voltadas para a gestão de dados (SAYÃO; SALES, 2020).

Na presente proposta de modelo, consideramos cinco instâncias de infraestruturas necessárias: de padronização, tecnológicas, informacionais, profissionais e organizacionais.

- *Normas e padrões* – Normas e padrões são formas consensuais de codificar o conhecimento que circula transversalmente por comunidades para assegurar uniformidade e similitude nos seus produtos e processos através do tempo e do espaço. Eles refletem o conhecimento mais atual sobre as práticas profissionais e aumentam a interoperabilidade, a consistência, a preservação, a reusabilidade, a segurança e a proteção das coleções digitais. Portanto, assegurar que em um ecossistema científico, em que as infraestruturas estão globalmente dispersas, seus produtos se alinhem aos Princípios FAIR, bem como tenham um grau satisfatório de qualidade e excelência e sejam apropriados às necessidades dos pesquisadores, exigem um corpo de padrões e princípios amplamente adotados e compartilhados. Considerando este fato, propõe-se que um corpo consensual de normas e padrões consubstanciem infraestruturas que devem estar subjacentes aos processos de gestão de dados. Isto porque, espera-se que as coleções de dados estejam aptas para serem utilizadas para uma grande variedade de propósitos – e não somente para as finalidades para as quais elas foram inicialmente coletadas. Para tal, elas precisam ser agregadas a outras coleções em outros sistemas, compartilhadas, acessadas, analisadas e arquivadas usando um amplo espectro de tecnologias. Essa condição torna um corpo de normas e padrões comuns infraestrutura essencial para a gestão e curadoria de dados de pesquisa. À medida que os princípios e práticas da gestão de dados de pesquisa se desenvolvem, eles começam a adquirir reconhecimento como um campo de conhecimento distinto e chamando a atenção de organizações interessadas no seu aprimoramento como, por exemplo, DCC, Codata, GOFAIR, DataOne, DataCite, entre muitas outras. Nesta direção, padrões e normas comumente adotadas para gestão de dados estão tomando corpo em muitas disciplinas e setores diferentes e estão sendo redefinidas em outras disciplinas. Como resultado, práticas aprimoradas para garantir a qualidade e a durabilidade dos dados digitais estão sendo continuamente estabelecidas. (NATIONAL RESEARCH COUNCIL, 2015).
- *Infraestrutura Tecnológica* – Compreende um vasto conjunto de atividades, equipamentos, processos e expertises que possam viabilizar os requisitos tecnológicos operacionais necessários às ciberinfraestruturas de gestão de dados, tais como: organização lógica, física e virtual dos dados; dispositivos para processamento de alto desempenho, computação em grade e armazenamento das coleções de dados locais ou em nuvem; redes locais, comunicações, conexões externas, internet, serviços web; aquisição/desenvolvimento de códigos científicos, software de *workflow*; equipamentos para análise de

dados e visualização; estratégias de segurança física, lógica e de rede.

- *Infraestrutura Informacional* – Compreende esquemas de representação e identificação persistentes; metadados descritivos, técnicos, administrativos, de preservação e disciplinares; além de taxonomia, ontologias, esquemas de classificação; bases de dados; inclui ainda repositórios, bibliotecas digitais e plataformas confiáveis para arquivamento de longo prazo.
- *Infraestrutura de pessoal* – As inúmeras instituições de pesquisa desenvolvem os mais diversos enfoques de gestão de dados. Isto pressupõe equipes de apoio compostas por diferentes profissionais (PINFIELD; COX; SMITH, 2014). Papéis como administrador de dados e cientistas de dados estão emergindo no mundo da ciência contemporânea e se incorporando às equipes mais tradicionais compostas por pesquisadores, técnicos de laboratório, assistentes de pesquisa e analistas; por outro lado, no âmbito das bibliotecas especializadas e dos repositórios, novos *stakeholds* como bibliotecários e arquivistas de dados e curadores, fazem a conexão entre a biblioteca e os laboratórios e apoiam a gestão das idiossincrasias disciplinares dos ciclos de vida dos dados (BALL, 2012). Entretanto, um requisito essencial – especialmente quando se trata dos serviços associados à curadoria – é a necessidade de conhecimento das disciplinas e domínios nos quais os dados são coletados, processados e utilizados. Sem alguma familiaridade com o problema a ser abordado, a cultura disciplinar, os objetivos a serem perseguidos, bem como os métodos utilizados, nomenclatura e práticas dos campos em que os ativos digitais são usados, os curadores não serão capazes de tomar as decisões mais corretas para gerenciarem esses ativos para uso atual e futuro (NATIONAL RESEARCH COUNCIL, 2015).
- *Infraestrutura organizacional* – O arcabouço infraestrutural pressupõe, assim como a governança, uma ancoragem baseada em alguma estrutura organizacional voltada para a pesquisa, como uma universidade, instituto de pesquisa, ou mesmo uma empresa cujos empreendimentos dependem da gestão de dados. Estas organizações precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

Essas vertentes infraestruturais – que possibilitam imbricamento de saberes e práticas que estão subjacentes a equipamentos, instalações, metodologias e principalmente a pessoas – proporcionam uma vasta carteira de serviços, ferramentas e processos que continuamente levam os objetos de pesquisa para um alinhamento com os Princípios FAIR. Nem sempre esses limites são claros, os repositórios, por

exemplo, são pontos de agregação de tecnologias, padrões, recursos informacionais e expertise em torno do arquivamento de objetos de pesquisa e constituem um elo imprescindível para se alcançar a Internet de Dados e Serviços FAIR, aproximam os vários estágios do ciclo de vida da gestão de dados aos ciclos de vida dos dados em seus ambientes de pesquisa.

5. Serviços para FAIRificação de dados

Para começar, é preciso esclarecer que tratamos aqui de serviços oferecidos pelas diversas plataformas de gestão de dados para dotar os objetos de pesquisa de graus de alinhamento aos Princípios FAIR. Estes serviços são de natureza distinta dos serviços oferecidos pela IFDS aos seres humanos e agentes computacionais, para benefícios de pesquisadores e outros *stakeholders*. Assim sendo, os serviços para FAIRificação podem ser classificados como informacionais, computacionais e científicos:

- *Serviços informacionais* – Compreendem os serviços oferecidos pelos profissionais de informação no âmbito de organizações como bibliotecas científicas e centros de informação: identificação persistente de objetos de pesquisa e pesquisadores; desenvolvimento de estruturas de representação como esquemas de metadados, taxonomia e ontologias; catalogação e indexação de objetos de pesquisa; publicação de dados; divulgação; letramento de pesquisadores; desenvolvimento de coleções de dados; apoio à elaboração de planos de gestão de dados; arquivamento de longo prazo/preservação; linking/contextualizaçãp.
- *Serviços computacionais* – Compreende disponibilidade de ferramentas de *software* e recursos de computação para apoiar o processamento, análise e visualização dos dados de pesquisa; recomendar como os dados podem melhor ser estruturados e armazenados e trabalhar, se necessário, junto aos pesquisadores na estruturação de bases de dados e marcação de texto (Wilson *et al*, 2011); estes serviços podem incluir ainda treinamento específico para a equipe de pesquisadores nos recursos oferecidos e em situações mais avançadas, oferecer processamento de alto desempenho e computação em grade.
- *Serviços científicos* – Compreendem os serviços que estão circunscritos ao ambiente científico, como laboratórios, e executados por pesquisadores ou especialistas em gestão de dados com conhecimentos disciplinares. São serviços relacionados à preparação de dados para usos mais amplos e podem incluir atividades como, avaliação, limpeza, normalização, organização dos arquivos, nomeação e, quando necessário, anonimização e outras estratégias

para preservação da privacidade, indexação disciplinar; documentação de códigos, *workflow* e processamento, agregação de dados. Mesmo considerando que esses serviços são protagonizados pelos próprios pesquisadores, eles precisam de considerável suporte computacional.

Os serviços que apoiam os processos de FAIRificação, na direção de uma Internet de Dados & Serviços FAIR, têm como ponto focal alguns conceitos que são essenciais para a concretização de seus pressupostos de reuso. São eles: acionalidade por máquina; metadados; e condições de acesso.

- *FAIR É SOBRE ACIONALIDADE POR MÁQUINA* – "O reconhecimento de que os computadores devem ser capazes de acessar dados publicados de forma autônoma, sem ajuda de operadores humanos é central para os Princípios FAIR", afirmam de forma categórica Mons e seus colaboradores (2017, p.51); assim sendo, "os princípios FAIR colocam uma ênfase privilegiada no aprimoramento das potencialidades das máquinas em encontrar e usar os dados, além de apoiar seu reuso por seres humanos" ratificam Wilkinson e colaboradores (2016, p.1). Isto fica claro quando se observa que grande parte do ciclo de vida dos dados, como indexação, recuperação via API, processamento e análise confiável de dados sensoriais são procedimentos assistidos e executados por computador, colocando em destaque o conceito de "acionável por máquina". De forma geral, este conceito pressupõe um contínuo de possíveis estados em que um objeto digital fornece informações cada vez mais detalhadas para um explorador de dados computacionais de ação autônoma. Os "*stakeholders* computacionais", como os denominou WILKINSON *et al*, 2016, tais como programas de aplicação e agentes computacionais, são exploradores que agem em nosso nome – seres humanos -, performando um papel crescentemente relevante na recuperação e análise de dados. Nesse contexto em constante transição, é preciso, portanto, considerar que os seres humanos não são os únicos interlocutores críticos no ecossistema de dados. Os Princípios FAIR são também, e principalmente, para máquinas.

Considerando a limitação primária dos seres humanos de operar no escopo, escala e velocidade requisitada pelo nível de complexidade da pesquisa contemporânea, especialmente no escopo da eScience, fica patente a necessidade das máquinas serem capazes de agir de forma autônoma e apropriada quando estiverem diante do amplo espectro de tipos, formatos, protocolos e mecanismos de acesso encontrados na exploração do ecossistema global de dados. "Um dos grandes desa-

fios da ciência intensiva de dados é, portanto, aprimorar a descoberta de conhecimento por meio da assistência de seres humanos e de seus agentes computacionais" ratificam (WILKINSON *et al*, 2016, p. 3). Esta interlocução é de grande importância na recuperação, acesso, integração e para "os tipos de análises integrativas profundas e amplas que constituem a maior parte da eScience contemporânea" (WILKINSON *et al*, 2016, p. 3).

Essas configurações e condições da ciência atual têm um reflexo profundo nos processos das modernas plataformas de gestão de dados, e a adoção total ou parcial dos princípios FAIR como parte da espinha dorsal desses sistemas técnicos-gerenciais é um passo importante na direção da acionabilidade por máquina, na medida em que as habilita a otimizarem o uso dos recursos de dados por meio de escolhas de implementação técnicas adequadas. Por exemplo, o recurso digital pode ser usado como um agente ou um substrato em análises baseado em aprendizagem por máquina ou inteligência artificial.

Por fim, é preciso observar que nem todos os dados podem ou devem obedecer a condição de serem automaticamente processados. Há inúmeras circunstâncias que, tornar os dados acionáveis por máquina, reduz a sua utilidade – por exemplo, quando falta ferramentas adequadas capazes de processar de forma eficiente determinados formatos (MONS *et al*, 2017).

- FAIR É SOBRE METADADOS – Fazendo uma ponte imprescindível entre acionabilidade por máquina e metadados Wilkinson e seus colaboradores (2016) nos esclarecem que um recurso, que se encontra num contínuo de possível estado acionável por máquina, fornece informações cada vez mais detalhadas a um explorador computacional, e isto se aplica em dois principais contextos: primeiro, se referindo aos metadados contextuais que envolvem o objeto digital, ou seja, reconhecendo o que é o objeto digital; segundo, quando se referindo ao conteúdo do objeto digital propriamente dito (como processá-lo/integrá-lo?). Nessa direção, essas informações – dependendo da quantidade, estruturação e qualidade – permitem a um agente que está diante de um objeto digital não encontrado anteriormente: identificar o tipo de objeto em relação à estrutura e intenção; identificar a sua utilidade no contexto considerado; determinar se ele pode ser usado de acordo com sua licença, consentimento, nível de sensibilidade ou limites de uso; e proceder ações apropriadas, tal como um ser humano faria. Assim, assistir as máquinas na descoberta e exploração de dados por meio de aplicações de tecnologias e padrões no nível das plataformas de dados se torna a prioridade máxima de uma boa gestão de dados e coloca em relevo a essencialidade do conceito de

metadados. Os padrões de metadados cumprem um papel-chave no fluxo da comunicação científica, cuja ênfase estende os requisitos metodológicos e de transparência do relato científico para o domínio da gestão de dados. Assim sendo, os Princípios FAIR enfatizam a importância dos metadados e de seus padrões na gestão de dados, focalizando o conceito de "metadado" transversalmente nos seus 15 princípios orientadores. "A mensagem-chave dos Princípios FAIR é que metadados e padrões de metadados devem ser articulados e tornados publicamente disponíveis na maior amplitude possível" (BOECKHOUT; ZIELHUIS; BREDENOORD, 2018, p. 932).

- FAIR É SOBRE ACESSO SOB CONDIÇÕES BEM DEFINIDAS – "FAIR não é igual a aberto", afirmam assertivamente JACOBSEN et al. (2020). O "A" no contexto do FAIR é compreendido como "Acessível sob condições bem definidas", o que o torna diferente de aberto sem restrições. Mons e colaboradores (2017, p.51) destacam que podem existir razões legitimas para blindar dados e serviços gerados com fundos públicos do acesso indiscriminado. Esses tipos de dados incluem: dados pessoais sensíveis, dados sobre geolocalização de espécie em perigo de extinção, sobre processos patenteáveis, segurança nacional, entre muitos outros. Além do mais, diversos setores, como o industrial e o médico, por razões legais, éticas, contratuais ou de competitividade, precisam de segurança apropriada para seus dados e requerem medidas adicionais de autorização e autenticação, tanto para exploradores humanos como para agentes computacionais; na prática, a Internet de Dados & Serviços FAIR não podem funcionar sem esses mecanismos (JACOBSEN *et al*, 2020). Embora mantenha conexões primordiais com os pressupostos da Ciência Aberta, os Princípios FAIR, explicita e deliberadamente, não endereçam questões éticas e morais sobre o grau de abertura dos dados, sua disponibilidade está inteiramente sob critério do custodiante dos dados. Os Princípios FAIR abordam apenas a necessidade de descrever um processo – automático ou manual – para acessar o dado descoberto; uma exigência de descrever de forma extensa e aberta o contexto no qual esses dados foram gerados.

Os princípios não necessitam que os dados FAIR sejam "abertos" ou "livres", entretanto eles exigem clareza e transparência acerca das condições que governam o seu acesso e reuso; requerem também que os dados tenham uma licença acessível e clara, preferencialmente legível por máquina. "O acesso transparente, porém, controlado dos dados e serviços, em oposição ao conceito genérico e ambíguo de "aberto", permite a participação de uma grande faixa de setores – públicos e privados - [...] ao redor do mundo", concluíram Mons *et al* (2017, p.52).

6. "FAIRificação" na direção da IFDS

A ideia primordial de implementação de uma Internet de Dados e Serviços FAIR não se realiza por si só. Para tal, é necessário um processo multidimensional de gestão de dados que possa efetivamente ir agregando valor, ao longo do tempo, aos objetos de pesquisa; o nível de aderência dos produtos de pesquisa aos Princípios FAIR está vinculado ao alcance e profundidade da gestão a que eles estão submetidos. Isto pressupõe a necessidade de um arcabouço de várias camadas – científica, tecnológica, informacional e de governança, conforme apresentado nas seções anteriores, que enderecem os inúmeros problemas éticos, metodológicos e organizacionais que se interpõem entre os fluxos de compartilhamento, integridade, reprodutibilidade, prestação de contas da pesquisa, bem como as novas necessidades e oportunidades de análise e reanálise em larga escala (WILKINSON *et al*, 2016, p.1).

No intuito de esclarecer os significados embricados no acrônimo FAIR, Mons e seus colaboradores (2017) oferecem uma escala de FAIRificação– aqui compreendida como o nível de profundidade e abrangência da gestão que tornem os objetos digitais de pesquisa aderentes aos Princípios FAIR. Nesse percurso, no grau mais baixo dessa escala estão os objetos sem nenhum potencial de reuso, que correspondem aos dados não publicados, ou publicados em ambientes instáveis como uma página web. Estes objetos não possuem **identificadores persistentes resolvíveis por máquina** que conduzam tanto aos elementos de dados quanto aos metadados correspondentes; estes, por sua vez, não são legíveis por máquina. O percurso mínimo em direção ao FAIRificação consiste em atribuir a um *dataset* um identificador persistente.

Porém, sem um conjunto de **metadados legíveis por máquina** será difícil encontrar o recurso, a menos que se conheça, a priori, o seu identificador. Isto indica que o identificador é necessário, porém insuficiente, e que é preciso ir mais adiante. O passo seguinte é a atribuição de metadados, que podem ter duas origens: "metadados intrínsecos", que são assinalados no momento da captura dos dados, geralmente por processos automatizados realizados pelos instrumentos ou *workflow* que geraram os dados, por exemplo, formato de arquivo, selo de tempo e localização; metadados assinalados pelos pesquisadores que criaram/coletaram os dados, profissionais de informação e os *stakeholds* que o reusaram na forma, por exemplo, de anotações, que conferem proveniência e contextualização aos dados e aumentam seu grau de FAIRificação. Por conseguinte, a adição de metadados ricos – e também FAIR – é um passo essencial nesse percurso. Assim sendo "a identificação persistente e a agregação de metadados já atribui um profundo efeito no potencial de reuso dos objetos de pesquisa, posto que podem ser identificados e recuperados". (MONS *et al,* 2017).

Contudo, mesmo que o dado seja tecnicamente FAIR ele pode estar com o acesso restrito por razões claras e justas tais como contratos, proteção de espécies em extinção, questões legais e éticas; isto dito, compreendemos que o padrão máximo de FAIRificação deve acontecer quando os próprios elementos de dados estão disponíveis sob condições bem definidas, para o reuso aberto por parte de qualquer outro interessado.

Indo ainda mais além na escala de FAIRificação, Barend Mons e seus colaboradores (2017) propõem que quando os dados estiverem linkados a outros objetos de pesquisa FAIR teremos alcançado a "Internet de Dados FAIR"; uma vez que um número crescente de aplicações e serviços podem linkar e processar dados FAIR, pode-se dizer que terá sido alcançada a "Internet de Dados & Serviços FAIR", significando um "ambiente global e compartilhado voltado para pesquisas orientadas por dados e inovação" (SALES *et al*, 2020, p.3), onde todos os pesquisadores podem acessar, armazenar, analisar e reusar dados para a pesquisa, inovação e para propósitos educacionais. A partir dos contornos desse território se estabelece uma ecologia de dados ativados por serviços associados que, para os diversos segmentos de usuários, se traduz num contínuo de benefícios acionados por aplicações computacionais.

Assim como a internet atual, que não tem uma governança centralizada e se baseia em um conjunto mínimo, porém rigoroso de padrões e protocolos que dão suporte a uma imensa variedade de implementações, o conceito de "Internet de Dados e Serviços FAIR" tem como pressuposto a máxima liberdade de desenvolvimentos por parte de todos os interessados. Nesse sentido, o roteamento escalável e transparente de DADOS, FERRAMENTAS e COMPUTAÇÃO – que processa (executa) as ferramentas – é a característica central de uma desejada Internet de Dados & Serviços, onde todos os tipos de provedores de serviços, públicos e privados, podem começar a implementação de protótipos de aplicações de dados e serviços FAIR (GO FAIR, [20--?a]).

Como abstração, a IFDS se modela na forma de hélice de três pás que correspondem aos elementos fundamentais – dados, ferramentas e computação - que são "roteados" para encontrar um ao outro no momento e no lugar certos e para serem usados e reusados da forma mais eficaz. Neste contexto, as ferramentas são principalmente definidas como serviços de software que agem sobre os dados, como por exemplo, máquinas virtuais empacotadas para viajar pelo IFDS fazendo análises distribuídas de dados ou mesmo um repositório de dados e a computação como a infraestrutura que capacita a ação. Assim como no modelo da ampulheta da internet, o eixo da hélice corresponde ao conjunto mínimo de padrões e protocolos, posto que o crescimento da ISDF está baseado no mantra da rede GO

FAIR: "Unicamente um conjunto mínimo necessário de protocolos e padrões para dar suporte a uma ampla variedade de escolhas de implementação para dados, ferramentas e elementos de computação". Os IFDS funcionariam de forma mais fluente se a infraestrutura subjacente operasse sobre uma rede forte, comum e globalmente interoperável e um motor que roteasse eficientemente dados para ferramentas, ferramentas para dados e ambos para a computação necessária, posto que esses três elementos cada vez mais não residem em grandes super sistemas de *storage* e HPC *facilities*, mas estão distribuídos por toda a internet (GO FAIR, [20--?a]; GO FAIR, [20--?b]).

7. À guisa de conclusão

A ciência contemporânea, intensiva em dados por natureza, exige um gestão de dados cuja escala extrapola as medidas mais convencionais, e precisa continuamente colocar esses ativos e outros objetos de pesquisa prontos para o reuso – o objetivo finalístico da gestão - por seres humanos e, sobretudo, por provedores de serviços, por meio de aplicações computacionais, ampliando, dessa forma, o seu potencial de reuso, de repropósito e de ressignificação para vários segmentos, incluindo os que estão fora do mundo da pesquisa. As dificuldades dos agentes humanos operarem na escala e velocidade exigidas pela complexidade das ciências intensivas de dados, especialmente a eScience, reforçam a necessidade de exploradores computacionais agirem de forma autônoma e apropriada em face de um ecossistema global de dados.

Entretanto, para alcançar esse estado de contínua oferta é necessária uma cadeia de processos que vão do estabelecimento de políticas a um alto grau de padronização que requisita um arcabouço infraestrutural cuja densidade depende do nível e profundidade da gestão. Mas o que se constata é que esse esforço, por vezes entrópico e com objetivos difusos, precisa de uma ordenação e de um horizonte. A aplicação dos princípios FAIR realinha esses esforços e estabelece objetivos claros para a gestão de objetos de pesquisa sintetizados nos seus quatro princípios fundamentais dimensionados pelos seus quinze princípios orientadores.

Nessa ecologia complexa, o modelo procurou desconstruir os blocos de construção (como legos) que compõem uma arquitetura genérica para se alcançar um nível de FARIficação que permita o alcance da almejada IFDS articulando os vários módulos conceituais – diretrizes, políticas, serviços, ferramentas, infraestruturas etc. – na forma de peças que podem ser ajustadas de acordo com a profundidade, alcance e filosofia de cada instituição ou disciplina, proporcionando, dessa forma, uma possível escala para apoiar a mensuração do nível de maturidade dos projetos de serviços de gestão.

Mesmo tendo em conta o enfoque generalista do modelo, é preciso considerar que na implantação de práticas e infraestruturas FAIR o contexto específico das comunidades científicas e as possibilidades da adoção devem ser observadas. A importância de cada princípio pode depender das prioridades e maturidade da comunidade e da geração e uso de determinados objetos de pesquisa. Essa condição implica que diferentes disciplinas encontrem soluções técnicas e necessitem de arcabouços infraestruturais e organizacionais e serviços de gestão diferentes para alcançar o grau de FAIRificação requeridos por suas comunidades. Mas é preciso observar que embora os imperativos científicos sejam diferentes entre disciplinas - que ainda apresentam diferentes tipos de organização e de cultura -, que faz com que elas busquem soluções próprias e sigam estratégias particulares na direção dos dados FAIR, as dificuldades e desafios, bem como as facilidades, são geralmente compartilhados, posto que há um núcleo comum de interesse. Além do mais, quando se amplia o alcance dos princípios FAIR para incluir outros objetos de pesquisa, é preciso considerar que muitos desses objetos pertencem a um domínio disciplinar específico, o que reforça a constatação de que as orientações e práticas FAIR são também específicas de disciplinas.

8. Referências

BALL, Alex. **Review of data management lifecycle models**. Bath, UK: University of Bath, 2012. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4219&rep=rep1&type=pdf. Acesso em: 25 abr. 2021.

BOECKHOUT, Martin; ZIELHUIS, Gerhard A.; BREDENOORD, Annelien L. The FAIR guiding principles for data stewardship: fair enough?. **European Journal of Human Genetics**, v. 26, n. 7, p. 931-936, 2018. Disponível em: https://www.nature.com/articles/s41431-018-0160-0.pdf. Acesso em: 25 abr. 2021.

BORGMAN, Christine. **Scholarship in the Digital Age**: Information, Infrastructure, and the Internet. London: The MIT Press, 2007.

COATES, Heather L. Building Data Services from the Ground Up: Strategies and Resources. Journal of eScience Librarianship, *v. 3, n.1,* 2014. Disponível em: https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1063&context=jeslib. Acesso em:25 abr. 2021.

ERWAY, Ricky *et al*. **Building Blocks:** Laying the Foundation for a Research Data Management Program. Dublin: OCLC, 2016. Disponível em: https://files.eric.ed.gov/fulltext/ED589141.pdf. Acesso em: 25 abr. 2021.

GO FAIR. GO FAIR **Initiative**. [20--?b]. Disponível em: https://www.go-fair.org/go-fair-initiative/. Acesso em: 25 abr. 2021.

GO FAIR. **The internet of FAIR Data & Service**. [20--?a]. Disponível em: https://www.go-fair.org/resources/internet-fair-data-services/. Acesso em: 25 abr. 2021.

GRAAF, Maurits van der; WAAIJERS, Leo. **A surfboard for riding the wave**: Towards a four country action programme on research data. Copenhagen: Knowledge Exchange, 2011. Disponível em: https://www.voced.edu.au/content/ngv%3A48428. Acesso em: 25 abr. 2021

HONG, Neil Chue *et al*. **Six recommendation to implementation of FAIR Practices**. Bruxelas: European Commission, 2020. Disponível em https://ec.europa.eu/info/publications/six-recommendations-implementation-fair-practice_en. Acesso em: 25 abr. 2021.

JACOBSEN, Annika *et al*. FAIR principles: Interpretations and implementation considerations. **Data Intelligence**, n. 2, p. 10–29, 2020. Disponível emhttp://www.inf.ufes.br/~gguizzardi/102-Annika_Jacobsen-1_GRFHSzW.pdf. Acesso em: 25 abr. 2021.

KOUPER, Inna *et al.* Research Data Services Maturity in Academic Libraries. *In*: JOHNSTON, Lisa R. (ed.). Curating Research Data*: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries, 2017.* p. 153-170. Disponível em: https://experts.illinois.edu/en/publications/research-data-services-maturity-in-academic-libraries. Acesso em: 25 abr. 2021.

LEONARDI, Paul M. Digital materiality? How artifacts without matter, matter. **First Monday**, v. 15, n 6 – 7, 2010. Disponível em: https://journals.uic.edu/ojs/index.php/fm/article/view/3036. Acesso em: 25 abr. 2021.

MAYERMIK, Mathews S. *et al*. The data conservancy instance: infrastructure and organizational services for research data curation. **D-Lib Magazine**, v.18, n.9/10, Sep./Out. 2012. Disponível em: http://www.dlib.org/dlib/september12/mayernik/09mayernik.html. Acesso em 25 abr. 2021.

MONS, Barend *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, v. 37, n. 1, p. 49-56, 2017.

MUSHI, G.E., PIENAAR, H., van DEVENTER, M. 2020. Identifying and Implementing Relevant Research Data Management Services for the Library at the University of Dodoma, Tanzania. **Data Science Journal**, v.19, n. 1, p. 1–9, 2020. Disponível em: Acehttps://datascience.codata.org/articles/10.5334/dsj-2020-001/. Acesso em 25 abr. 2021.

NATIONAL RESEARCH COUNCIL. **Preparing the workforce for digital curation**. Washington, D.C.: The National Academies Press, 2015.

PINFIELD, Stephen; COX, Andrew M.; SMITH, Jen. Research data management

and libraries: Relationships, activities, drivers and influences. **PLoS One**, v. 9, n. 12, p. e114734, 2014. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114734. Acesso em: 25 abr. 2021.

REED, Robyn B. Diving into data: Planning a research data management event. **Journal of Escience Librarianship**, v. 4, n. 1, 2015. Disponível em : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517608/. Acesso em: 25 abr. 2021.

SALES, Luana *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/238/10004/GO-FAIR-Brazil-A-Challenge-for-Brazilian-Data. Acesso em: 25 abr. 2021. #31

SAYÃO, Luís Fernando. Modelos teóricos em ciência da informação-abstração e método científico. **Ciência da informação**, Brasília, v. 30, n. 1, p. 82-91, 2001. Disponível em:. Acesso em: 25 abr. 2021.

SAYÃO, L. F.; SALES, L. F. AFINAL, O que é dado de pesquisa? BIBLOS, v. 34, n. 2, 2020. Disponível em: https://www.seer.furg.br/biblos/article/view/11875. Acesso em: 12 maio. 2021.

SOLOMONIDES, Anthony. Research Data Governance, Roles, and Infrastructure. *In*: RICHESSON, Rachel; ANDREWS, James (eds.). **Clinical Research Informatics**. Cham: Springer, 2019. p. 291-310.

STRASSER, Carly. **Research data management**. Baltimore: NISO, 2015. Disponível em: https://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf. Acesso em: 25 abr. 2021.

TENOPIR, C., BIRCH, B., ALLARD, S. **Academic libraries and research data services**: Current practices and plans for the future. An ACRL White Paper. Chicago, IL: Association of College and Research Libraries, 2012.Disponível em: https://trace.tennessee.edu/utk_dataone/20/. Acesso em: 25 abril 2021.

WILKINSON, Mark D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016. Disponível em: https://www.nature.com/articles/sdata201618.pdf. Acesso em: 25 abr. 2021.

WILSON, James A. J. *et al*. An institutional approach to developing research data management infrastructure. **The International Journal of Digital Curation**, v. 6, n. 2, 2011. Disponível em: http://ijdc.net/index.php/ijdc/article/view/198. Acesso em: 25 abr. 2021.

Como citar com o DOI individual

SAYÃO, Luís Fernando; SALES, Luana Farias. Um modelo de implementação para a internet de dados & serviços FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 215 - 242. DOI: 10.22477/9786589167242.cap16

Seção 5

PRÁTICAS DE GESTÃO DE DADOS FAIR NO AMBITO DO GO FAIR BRASIL

Rede GO FAIR Brasil Saúde Enfermagem: onde estamos e aonde queremos chegar?

Eliza Macedo[1], Patrícia Henning[2], Maria Simone de Menezes Alencar[3] e Sônia Souza[4]

1. Introdução

Os debates sobre a produção do conhecimento e o trabalho científico, bem como as reflexões sobre as relações entre ciência, tecnologia, dados e informação são temas em constante evolução desde o século passado. Nos dias atuais, os referidos temas vêm adotando novas configurações nas formas de avaliação, de gestão, de disseminação e de armazenamento moldando-se às tendências do mundo digital, voltadas para o conhecimento comum, coletivo e colaborativo.

É nesse contexto que um novo paradigma vem se configurando no campo científico, corroborando com as práticas científicas abertas, que integram atores tecnológicos e humanos, direcionadas para o coletivo, estimulando o compartilhamento, o reuso e a preservação digital dos dados. Impondo-se, no entanto, a necessidade de se repensar novas diretrizes e políticas que possam atender melhor as demandas dessa nova realidade voltadas para a Ciência Aberta, considerada um fenômeno internacional, fruto das tendências de democratização do saber.

1 Doutora em Enfermagem e Biociências, Professor Associado iv da Escola de Enfermagem Alfredo Pinto. Universidade Federal do Estado do Rio de Janeiro (unirio), eliza.macedo@unirio.br

2 Doutora em Comunicação e Informação em Saúde (ppgics/Fiocruz), Professora Visitante Programa de Pós-Graduação em Enfermagem (ppgenf) da Universidade Federal do Estado do Rio de Janeiro (unirio), henningpatricia@gmail.com

3 Doutora na área de Gestão e Inovação Tecnológica pela Escola de Química da ufrj, Professora permanente do Mestrado Profissional em Biblioteconomia (ppgb) e do Doutorado em Enfermagem e Biociências (ppgenfbio) da Universidade Federal do Estado do Rio de Janeiro (unirio), simone.alencar@unirio.br

4 Doutora em Enfermagem. Professor Associado iv da Escola de Enfermagem Alfredo Pinto. Universidade Federal do Estado do Rio de Janeiro (unirio), sonia.souza@unirio.br

Dentre todas as práticas da Ciência Aberta, os dados abertos de pesquisa são considerados insumos do trabalho científico que obtiveram maior destaque e importância nos dias de hoje, devido à necessidade e urgência do seu compartilhamento, tão logo sejam gerados, em áreas estratégicas como a da saúde, visando o reuso sempre que possível. Essa possibilidade proporciona maior rapidez, transparência e agilidade às pesquisas alavancando a produção de conhecimento e da ciência.

No entanto, devido à complexidade dos dados e às especificidades de cada área do conhecimento, aumenta a necessidade de contextualização e organização dos dados, assim como o detalhamento da sua proveniência, para garantir a sua preservação a longo prazo e o reuso em outras pesquisas. É dentro desse contexto que surgem os princípios FAIR, considerados internacionalmente como norteadores das boas práticas de gestão dos dados de pesquisa e de iniciativas voltadas para a sua disseminação e implementação.

Este relato tem por objetivo situar a Rede GO FAIR Brasil Saúde Enfermagem, descrever sua trajetória de implementação até o momento e articular os mecanismos de atuação dessa iniciativa internacional na área da Enfermagem. Para tanto, se faz necessário uma breve introdução acerca dos princípios FAIR, da iniciativa internacional GO FAIR, do escritório GO FAIR no Brasil e da Rede GO FAIR Brasil Saúde, que constituem o conteúdo das próximas seções.

2 Os princípios FAIR e a iniciativa GO FAIR internacional

Em janeiro de 2014, surgiram as primeiras manifestações sobre as questões relacionadas à gestão de dados quando um grupo de especialistas, editores científicos, representantes da academia, de agências de fomento à pesquisa e da área industrial se reuniram em um workshop intitulado *Jointly designing a data* FAIRPORT, no Lorentz Centre, em Leiden, Holanda. Esse encontro foi marcado pelo alto nível de discussão em torno da criação de uma infraestrutura global que pudesse dar suporte às publicações, descobertas, compartilhamento e reutilização dos dados de pesquisa. Fruto desse encontro, foi elaborado um conjunto de diretrizes voltadas para as boas práticas de gestão de dados denominada "FAIR *Principles*". Tais princípios são na realidade um acrônimo para *Findable, Accessible, Interoperable and Reusable* (FAIR) e só foram oficialmente publicados em 2016, na revista *Scientific Data*, no artigo de Wilkinson *et al* (2016) intitulado *The* FAIR *Guiding Principles for scientific data management and stewarship.* A figura 1 apresenta, de forma sucinta, os princípios FAIR.

É possível observar que esses princípios, por si só, não trazem muitos esclarecimentos a respeito da sua implementação. Eles descrevem apenas um conjunto de atributos desejados para as boas práticas de gestão e tratamento dos recursos digitais. Mons *et al* (2017, p.50) esclarecem que eles "deliberadamente não especifi-

cam requisitos técnicos, mas sim um conjunto de orientações para uma crescente reutilização contínua, por meio de diferentes implementações".

Considerando que as pesquisas científicas estão sendo conduzidas por dados cada vez mais complexos, exigindo não apenas melhor tratamento e organização, mas também maior capacidade das máquinas, *software* apropriados, recursos humanos melhor capacitados e recursos financeiros elevados para lidar com tal realidade, surge em 2017 a iniciativa *Global Open* FAIR (GO FAIR)[5] com o intuito de disseminar os princípios e serviços FAIR e dar orientações basilares para a sua implementação nesse cenário da ciência de dados.

Figura 1 – Princípios FAIR

Fonte: WILKINSON *et al*, 2016. Adaptação e tradução das autoras.

Essa iniciativa adota uma abordagem *bottom-up* como metodologia de implementação, ou seja, incentiva a criação de redes independentes e autônomas criadas pela comunidade científica, por livre e espontânea vontade, seguindo sua filosofia e orientações.

3. A rede GO FAIR brasil saúde enfermagem: onde estamos?

O primeiro encontro da iniciativa GO FAIR Brasil aconteceu em 25 de setembro de 2018, em São Paulo, durante os 20 anos da Rede *Scientific Eletronic Library Online* (SciELO), onde ficou instituído o Instituto Brasileiro em Informação Científica e Tecnológica (IBICT) como o coordenador do escritório do GO FAIR no Brasil. Estiveram presentes diversos *stakeholders* representantes de universidades e insti-

5 Disponível em: https://www.go-fair.org/. Acesso em: 03 dez. 2020.

tutos de pesquisa brasileiros, bem como representantes da GO FAIR Internacional (SALES *et al*, 2020)

No entanto, o seu lançamento oficial, para a comunidade científica, ocorreu no mês seguinte, em 10 de dezembro de 2018, durante evento promovido pelo Ministério de Ciência, Tecnologia, Inovação e Comunicação (MCTIC). A GO FAIR Brasil[6] tem a responsabilidade de difundir, apoiar e coordenar as atividades de adoção de estratégias de implementação dos princípios FAIR, respeitando as especificidades das diferentes áreas do conhecimento, em todo o território nacional.

A rede GO FAIR Brasil Saúde[7] é a primeira rede de implementação brasileira considerada hoje a mais atuante, sendo responsável pela elaboração de estratégias de adoção dos princípios FAIR nos domínios da saúde. Sua coordenação está sob a responsabilidade do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), da Fundação Oswaldo Cruz (Fiocruz) e conta com a participação de diversas instituições das áreas de Saúde Pública, Vigilância Sanitária, Informação e Comunicação em Saúde, História do Patrimônio Cultural das Ciências e da Saúde, Oncologia, Enfermagem e Educação Profissional em Saúde.

Na área de Enfermagem, começaram a ser traçados os primeiros passos para a composição da sub-rede GO FAIR Brasil Saúde Enfermagem, considerada um dos braços da rede GO FAIR Brasil Saúde, em março de 2019, quando pesquisadores da Escola de Enfermagem Alfredo Pinto (EEAP), do Programa de Pós-Graduação em Saúde e Tecnologia do Espaço Hospitalar (PPGSTEH) (mestrado profissional), do Programa de Pós-Graduação em Enfermagem (PPGENF) (mestrado acadêmico) e do Programa de Pós-Graduação em Enfermagem e Biociências (PPGENFBio) (Doutorado) se reuniram pela primeira vez com representantes da iniciativa GO FAIR Internacional, do GO FAIR Brasil e da Rede GO FAIR Brasil Saúde, para pactuar a criação da Rede GO FAIR Brasil Saúde Enfermagem (HENNING, 2019).

Desde então, várias ações vêm sendo desenvolvidas para o seu fortalecimento e implantação. Primeiramente foi oferecido à comunidade de enfermagem brasileira o "Seminário Internacional sobre Gestão de Dados de Pesquisa em Saúde - GO FAIR Brasil Saúde e GO FAIR Brasil Saúde Enfermagem", que aconteceu durante todo o mês de junho de 2020 e que contou com a presença de mais de 250 participantes. A programação científica desse evento foi concebida utilizando-se da modalidade de *webinars* e proferida por palestrantes de instituições do Brasil e da Holanda, com reconhecida produção intelectual e trajetória sobre a temática.

6 Disponível em: https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office/. Acesso em: 03 dez. 2020.

7 Disponível em: https://portal.fiocruz.br/go-fair-brasil-saude. Acesso em: 03 dez. 2020.

O conteúdo programático desse Seminário foi planejado para contemplar todas as práticas relacionadas à gestão de dados de pesquisa e aos princípios FAIR, distribuído em oito módulos. Módulo 1: Introdução aos Dados de Pesquisa[8]; Módulo 2: Introdução aos Princípios FAIR e a Iniciativa GO FAIR[9]; Módulo 3: Plano de Gestão de Dados no contexto da COVID19[10]; Módulo 4: Repositórios de dados de pesquisa no contexto da COVID19[11]; Módulo 5: Preservação e Curadoria de dados[12]; Módulo 6: Interoperabilidade de dados[13]; Módulo 7: Tecnologias FAIR para reprodutibilidade de pesquisa[14]; Módulo 8: Projeto VODAN Brasil – Rede de dados de pesquisa para enfrentamento da COVID19.[15] Todos os vídeos das apresentações estão disponíveis no Repositório ARCA da Fiocruz e nas páginas do *YouTube* da Escola de Enfermagem Alfredo Pinto.[16]

Figura 2 – Flyer de divulgação do Seminário Internacional

Fonte: http://www.unirio.br/prae/ppgsteh/noticias-1/seminario-internacional-sobre-gestao-de-dados-de-pesquisa-em-saude-1

8 Disponível em: https://www.arca.fiocruz.br/handle/icict/45046. Acesso em: 03 dez. 2020.
9 Disponível em: https://www.arca.fiocruz.br/handle/icict/45049. Acesso em: 03 dez. 2020.
10 Disponível em: https://www.arca.fiocruz.br/handle/icict/45050. Acesso em: 03 dez. 2020.
11 Disponível em: https://www.arca.fiocruz.br/handle/icict/45052. Acesso em: 03 dez. 2020.
12 Disponível em: https://www.arca.fiocruz.br/handle/icict/45058. Acesso em: 03 dez. 2020.
13 Disponível em: https://www.arca.fiocruz.br/handle/icict/45059. Acesso em: 03 dez. 2020.
14 Disponível em: https://www.arca.fiocruz.br/handle/icict/45060. Acesso em: 03 dez. 2020.
15 Disponível em: https://www.arca.fiocruz.br/handle/icict/45061. Acesso em: 03 dez. 2020.
16 Disponível em: https://www.youtube.com/channel/UCH-mOJCskxwnHQweoPkNV-g. Acesso em: 03 dez. 2020.

É importante ressaltar que a Rede GO FAIR Brasil Saúde Enfermagem é coordenada pelo Programa de Pós-Graduação em Saúde e Tecnologia no Espaço Hospitalar (PPGSTEH), em gestão colegiada com a Escola de Enfermagem Alfredo Pinto (EEAP), pelo Programa de Pós-Graduação em Enfermagem e Biociências (PPGENFBIO) e pelo Programa de Pós-Graduação em Enfermagem (PPGENF), da Universidade Federal do Estado do Rio de Janeiro (UNIRIO).

Após deliberação colegiada com os coordenadores dos Programas de Pós-Graduação em Enfermagem da UNIRIO, formou-se um grupo de docentes que integram os programas com o objetivo de acompanhar e implementar as atividades de implantação da Rede GO FAIR Brasil Saúde Enfermagem. O grupo é formado pelos seguintes docentes: Drª Eliza Macedo, Drª Patrícia Henning, Drª Maria Simone de Menezes Alencar, Drª Sônia Souza, Drª Danielle Galdino, Drª Taís Vernaglia e Drª Inês Meneses. Dentre as ações do grupo podemos citar: reuniões para o planejamento estratégico de implementação; reunião com representante da FIOCRUZ para os trâmites do acordo de Cooperação Técnica; cadastro do Projeto de Pesquisa GO FAIR Brasil Saúde Enfermagem no Departamento de Pesquisa da UNIRIO; elaboração de projeto destinado a alunos de graduação interessados na temática, para atuarem como bolsistas, já enviado à Pró-reitoria de Assuntos Estudantis, que oferece bolsas de incentivo acadêmico; e cadastro das ações na Pró-reitoria de Extensão, ambas da UNIRIO.

Dois meses após a realização do Seminário Internacional sobre Gestão de Dados de Pesquisa em Saúde, ocorreu o lançamento da Rede GO FAIR Brasil Saúde Enfermagem, em 22 de setembro de 2020, durante as comemorações dos 130 anos do aniversário da Escola de Enfermagem Alfredo Pinto, da UNIRIO. A abertura do lançamento foi feita com a participação dos coordenadores da GO FAIR Brasil, GO FAIR Brasil Saúde e GO FAIR Brasil Saúde Enfermagem e com a divulgação do seu manifesto de adesão à Rede. Estiveram presentes 114 profissionais com perfis variados desde pesquisadores, docentes, discentes de pós-graduação e graduação, bibliotecários, arquivistas, técnicos administrativos à profissionais da área da saúde pertencentes a universidades, institutos de pesquisa e hospitais. Ao término das apresentações, os participantes foram convidados a preencher um cadastro e assinar o Manifesto Aberto de adesão à Rede GO FAIR Brasil Saúde Enfermagem,[17]

Os interessados em participar da Rede precisaram apenas assinar o Manifesto e ter disponibilidade de trabalhar de forma colaborativa. Os coordenadores irão entrar em contato e inserir os interessados nas dinâmicas de planejamento e ações futuras da Rede.

17 Disponível em: https://bit.ly/GOFAIRENFERMAGEM. Acesso em: 03 dez. 2020.

Em continuidade às ações, o Programa de Pós-Graduação em Saúde e Tecnologia no Espaço Hositalar (PPGSTEH), visando buscar novas adesões à Rede GO FAIR Brasil Saúde Enfermagem, ofereceu à comunidade científica de Enfermagem, nos dias 03 e 16 de dezembro de 2020, o workshop "A Rede GO FAIR Brasil Enfermagem: onde estamos e onde queremos chegar"[18]. Este evento teve o objetivo de apresentar à comunidade científica de enfermagem e aos interessados em atuar na área da saúde em geral as ações que vêm sendo desenvolvidas no âmbito da Rede GO FAIR Brasil Saúde Enfermagem, voltadas para a gestão de dados de pesquisa em enfermagem, no âmbito da Ciência Aberta.

Figura 3 – Flyer do Lançamento da Rede GO FAIR Brasil Saúde Enfermagem

Fonte: https://portal.fiocruz.br/noticia/seminario-virtual-marca-o-lancamento-da-rede-go-fair-brasil-saude-enfermagem.

Figura 4 – Flyers de divulgação do Workshop da Rede GO FAIR Brasil Saúde Enfermagem

Fonte: http://www.unirio.br/news/workshop-ira-discutir-gestao-de-dados-de-pesquisa-em-enfermagem.

18 Disponível em: http://www.unirio.br/news/workshop-ira-discutir-gestao-de-dados-de-pesquisa-em-enfermagem. Acesso em: 03 dez. 2020.

Conforme descrito em seu Manifesto, esta Rede se propõe a trabalhar no fortalecimento e disseminação dos princípios FAIR, no campo da enfermagem, de forma articulada e colaborativa com os seus membros. No momento, estão sendo traçadas as primeiras diretrizes de atuação da Rede, que se organizará por meio de grupos de trabalho específicos, com a participação e atuação voluntária da comunidade brasileira de Enfermagem.

4. A rede GO FAIR Brasil Saúde Enfermagem: aonde queremos chegar?

A Rede GO FAIR Brasil Saúde Enfermagem busca o seu desenvolvimento e consolidação por meio do fortalecimento e promoção da gestão, compartilhamento e reuso dos dados de pesquisa em Enfermagem, dentro dos Programas de Pós-Graduação em Enfermagem brasileiros, que são os principais geradores de dados de pesquisa em Enfermagem. Para isso tem como metas:

1) Promover pesquisas, na área de enfermagem, voltadas para metadados específicos, padrões de interoperabilidade tecnológica e semântica dos dados como: uso de vocabulários controlados e ontologias da área; modelos de plano de gestão de dados; repositórios de dados de pesquisa nacionais e internacionais, que poderão armazenar de forma confiável os dados de pesquisa em enfermagem; aplicação de licenças de uso, de acordo com os marcos regulatórios brasileiros.
2) Desenvolver metodologias voltadas para as práticas dos produtos e serviços FAIR, que atendam às necessidades disciplinares e operacionais da área da enfermagem.
3) Promover encontros, cursos, workshops e seminários visando impulsionar e disseminar os princípios FAIR entre os membros da Rede GO FAIR Brasil Saúde Enfermagem.
4) Criar grupos de trabalhos de forma voluntária e colaborativa, junto aos seus membros, que desenvolverão ações de capacitação, ações de desenvolvimento técnico e tecnológico e ações políticas voltadas para a expansão da Rede;
5) Trabalhar de forma articulada e colaborativa com a Rede GO FAIR Brasil Saúde, junto à Fiocruz.

Estas metas, criadas em reunião de coordenação da Rede, são de médio e longo prazo havendo a possibilidade de novas metas serem criadas ao longo do processo à medida que novas demandas comecem a surgir para a sustentabilidade da Rede. Reuniões quinzenais estão sendo programadas para o início das atividades em 2021

com o objetivo reunir os membros no desenvolvimento das atividades.

Considera-se que ainda há um longo caminho a percorrer até alcançar-se a tão almejada gestão de dados FAIR. Para isso acontecer, os programas de Pós-graduação em Enfermagem devem, paralelamente, integrar o conteúdo de gerenciamento de dados de pesquisa em seus currículos. Além disso, devem também seguir as sugestões de Raszewski *et al*. (2020, p. 7), que afirmam que ao criar uma infraestrutura de políticas, práticas e currículos que contemplem o gerenciamento de dados, formarão pesquisadores preparados para atender as expectativas de competência de dados de sistemas de saúde, clínicas de atenção primária, no âmbito da comunidade e da academia.

O interesse das comunidades de Enfermagem na implementação dos princípios FAIR fica evidente com a grande participação de profissionais da saúde em diversas atividades desenvolvidas em um curto período de tempo, tal como o Seminário Internacional em Gestão de Dados em Saúde e o Workshop da Rede GO FAIR Brasil Saúde Enfermagem. Isso indica um futuro promissor em nome da boa gestão de dados de pesquisa em enfermagem, voltada para estudos, desenvolvimento de infraestruturas e participação em fóruns nacionais e internacionais.

5. Considerações finais

É conhecido que vários tipos de dados são utilizados como insumo de pesquisa, desde dados governamentais, administrativos, de empresas privadas, científicos, assim como aqueles da área da saúde, onde destaca-se a Enfermagem. No âmbito da Ciência Aberta, os dados de pesquisa devem ser tratados e abertos tão logo quanto possível, respeitando as suas especificidades disciplinares e legais. Além dos obstáculos inerentes na abertura dos dados de pesquisa, como a falta de entendimento sobre as definições legais de propriedade intelectual e a falta de padronização nas definições e configurações dos dados de pesquisa, a área de saúde, dada as suas especificidades, tem um desafio a mais a pensar: a necessária proteção dos dados sensíveis, sendo este um ponto nevrálgico para os pesquisadores que ainda não têm familiaridade com o tema, sendo necessários *software* e treinamento adequados.

Diante deste cenário cheio de dúvidas e oportunidades, na construção de uma nova cultura que aborde todos os estágios do ciclo de vida dos dados, a Rede GO FAIR Brasil Saúde Enfermagem busca superar tais obstáculos, se apropriando do tema "Gestão de Dados FAIR", visando elaborar um conjunto de habilidades para orientar os alunos e pesquisadores dos Programas de Pós-Graduação em Enfermagem na criação de novos conteúdos com foco em dados e posteriormente, desenvolver infraestruturas de apoio ao desenvolvimento de planos de gestão de dados e

de armazenamento em repositórios de dados apropriados.

Por perceber a importância deste cenário que se amplia e se fortalece, o Programa de Pós-Graduação em Saúde e Tecnologia no Espaço Hospitalar (PPGSTEH), em gestão colegiada com a Escola de Enfermagem Alfredo Pinto, o Programa de Pós-Graduação em Enfermagem (PPGENF) (mestrado acadêmico) e o Programa de Pós-Graduação em Enfermagem e Biociências (PPGENFBio) (doutorado) dão início, por intermédio da coordenação da Rede GO FAIR Brasil Saúde Enfermagem, aos primeiros passos para a inserção dos dados de pesquisa em enfermagem no processo de compartilhamento e reuso dos dados alinhados aos princípios FAIR, que promoverá a cultura e trará impacto e visibilidade para os dados produzidos pela área da Enfermagem, articulando o agir local ao pensando global.

6. Referências

HENNING, Patrícia. Gestão de Dados de Pesquisa: uma demanda necessária para a geração de novos conhecimentos. Editorial. **Revista Online de Pesquisa:** Cuidar é Fundamental. v. 11, n. 3, 2019. Disponível em: http://www.seer.unirio.br/index.php/cuidadofundamental/article/view/8939/pdf. Acesso em: 03 dez. 2020.

MONS, Barend *et al*. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use,** v.37, n. 1, p. 49-66, 2017. Disponível em: https://content.iospress.com/articles/information-services-and-use/isu824. Acesso em: 03 dez. 2020.

RASZEWSKI, Rebecca *et al*. A survey of current practices in data management education in nursing doctoral programs. **Journal of Professional Nursing**. 2020. Disponível em: https://doi.org/10.1016/j.profnurs.2020.06.003. Acesso em: 03 dez. 2020.

SALES, Luana *et al.* GO FAIR Brazil: A Challenge for Brazilian Data Science. **Data Intelligence** 2:1-2, 238-245, 2020, Disponível em: https://doi.org/10.1162/dint_a_00046 Acesso em: 03 dez. 2020.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 1, p. 1-9, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 03 dez. 2020.

Esse projeto conta com o apoio da FAPERJ, por meio do fomento ao Laboratório de Ciência Aberta e Dados de Pesquisa para apoio à Inovação - LabINOVA.

Como citar com o DOI individual

MACEDO, Eliza; HENNING, ALENCAR, Patrícia Simone; Sônia Souza. Rede GO FAIR Brasil Saúde Enfermagem: onde estamos e aonde queremos chegar?. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 243 - 252. DOI: 10.22477/9786589167242.cap17

VODAN BR – uma plataforma de apoio para dados COVID-19 seguindo os princípios FAIR

Maria Luiza Machado Campos[1], Vania Borges[2], Giseli Rabello Lopes[3], Maria Claudia Cavalcanti[4], João Moreira[5], Sergio Manuel Serra da Cruz[6]

1. Introdução

A pandemia de COVID-19 deixou clara a importância de se ter os resultados de pesquisas científicas mais facilmente disponíveis para pronto e amplo reuso. Diversos grupos já envolvidos nesses temas se mobilizaram buscando discutir e agilizar a definição e construção de infraestruturas de apoio capazes de fazer frente a este desafio. Em especial, participantes da *Research Data Alliance* (RDA)[7], da World Data Systems (WDS)[8], da rede GO FAIR[9] e do *Committee on Data* (CODATA)[10], ligado ao *International Science Council* (ISC)[11], lançaram um apelo por ações, chamado *Data Together* (DATA TOGETHER, 2020), que acelerou e promoveu a cooperação entre diferentes iniciativas já em andamento.

A rede de implementação GO FAIR *Virus Outbreak Data Network* (VODAN) foi concebida para iniciar uma "comunidade de comunidades" para projetar e construir, de forma ágil, a infraestrutura para uma rede de dados internacional interoperável e distribuída, e para oferecer suporte na busca por respostas, baseadas em evidências, sobre casos de surto viral (MANISFESTO VODAN, 2020). Por ser uma iniciativa do consórcio GO FAIR, os dados e serviços gerados devem obedecer aos

1 PhD, Programa de Pós-Graduação em Informática – PPGI/UFRJ, mluiza.campos@gmail.com
2 Doutoranda, Programa de Pós-Graduação em Informática – PPGI/UFRJ, vjborges30@gmail.com
3 DSc, Programa de Pós-Graduação em Informática – PPGI/UFRJ, giseli@dcc.ufrj.br
7 https://www.rd-alliance.org/
8 https://www.worlddatasystem.org/
9 GO FAIR – iniciativa que incentiva a disponibilização de dados e serviços FAIR (Localizáveis, Acessíveis, Interoperáveis e Reutilizáveis) para projetos de pesquisas científicas https://www.go-fair.org
10 https://codata.org
11 https://council.science/

princípios FAIR (MONS, 2020), que fornecem diretrizes para tornar os dados localizáveis, acessíveis, interoperáveis e reutilizáveis. No caso da rede VODAN, o ponto de partida são os dados clínicos de pacientes com COVID-19, realizando-se, em uma primeira fase, a transformação e tratamento desses dados, de acordo com o Formulário de Pesquisa Clínica (FPC, - do original em inglês, *Clinical Research Form* – CRF), desenvolvido e padronizado pela Organização Mundial de Saúde (OMS).

O FPC da OMS (FPC-OMS) é um protocolo de pesquisas clínicas, desenvolvido com auxílio de especialistas, para captar informações consideradas relevantes em casos de epidemia e pandemia. Esse formulário encontra-se dividido em três módulos: o primeiro destinado a coletar os dados de admissão do paciente; o segundo para os dados referentes ao acompanhamento, durante a internação; e o terceiro para os dados do desfecho do tratamento, seja por alta, por transferência hospitalar ou por óbito.

De acordo com o manifesto VODAN (2020), a proposta original consiste no desenvolvimento de uma solução que permita aos profissionais de saúde registrarem os dados observados no formato estabelecido pelo FPC-OMS, armazenando-os em repositórios ou bancos de dados. Os metadados sobre esses repositórios devem ser, posteriormente, disponibilizados em um *FAIR Data Point* (FAIR DP). Um FAIR DP é um componente da infraestrutura de apoio a dados FAIR através do qual agentes de software podem ter acesso a descritores que permitem localizar e visitar os dados localmente e executar consultas sobre eles (MONS, 2020). O curador dos dados locais dará a permissão ou não para que a consulta/análise seja executada. Essa estrutura permite que as informações dos pacientes permaneçam protegidas nas bases de dados das unidades de saúde, respeitando a legislação para dados de saúde existente em cada país.

No Brasil, o projeto VODAN BR[12] teve início concomitantemente com o avanço da pandemia no país, ao longo dos primeiros meses de 2020, como parte da rede GO FAIR Brasil Saúde[13], vinculado à Fundação Oswaldo Cruz (FIOCRUZ), em parceria multi-institucional com a Universidade Federal do Rio de Janeiro (UFRJ) e a Universidade Federal do Estado do Rio de Janeiro (UNIRIO), entre outras instituições. O desenvolvimento da infraestrutura está sob a responsabilidade do Grupo de Pesquisa GRECO[14], da UFRJ, e tem como parceiros do teste piloto o Hospital Federal Gaffreé Guinle, do Rio de Janeiro, e o Hospital Municipal São José, de Duque de Caxias. Os dados são coletados de seus sistemas originais e tratados para se

12 https://vodanbr.github.io/

13 https://portal.fiocruz.br/go-fair-brasil-saude

14 http://dgp.cnpq.br/dgp/espelhogrupo/634046

alinharem ao padrão estabelecido, ou seja, ao formato de questionário da OMS, visando sua posterior disponibilização e de seus metadados, seguindo os princípios FAIR, padrões da Web Semântica e obedecendo critérios de licenciamento e anonimização estabelecidos.

Este capítulo tem por objetivo apresentar uma visão geral dos processos e ativo computacional sendo desenvolvidos para apoio ao projeto VODAN BR. Essa infraestrutura escalável, distribuída e genérica visa atender um processo intensivo de coleta de dados com alta heterogeneidade, disponibilizando-os em plataformas que ofereçam dados e metadados interoperáveis e processáveis por agentes de software, apoiando a descoberta de outros recursos que possam ser associados a eles. Com isso, é possível obter maior agilidade na descoberta e geração de conhecimento a partir de um reuso mais efetivo dos resultados de pesquisas.

As próximas seções estão estruturadas da seguinte forma: a seção 2 aborda a rede de implementação VODAN e tecnologias em que se apoia; a seção 3 apresenta a plataforma VODAN BR, descrevendo o processo e a infraestrutura sendo desenvolvidos; e a seção 4 apresenta a conclusão e possibilidades futuras identificadas.

2. A rede de implementação VODAN e tecnologias associadas

Apesar do volume de informações disponibilizadas na Web desde o início da pandemia ter crescido muito acima das expectativas, observa-se que, em sua grande maioria, elas se referem a totais de pessoas contagiadas, internadas, recuperadas e de óbitos. De forma complementar a esses dados agregados, dados referentes ao quadro clínico, ao tratamento dos pacientes e seu desfecho se constituem em importante apoio para estudos mais detalhados em pesquisa clínica. No entanto, de modo geral, observa-se que esses dados, apesar de serem extremamente valiosos para a comunidade científica, não se encontram, em geral, acessíveis.

Para lidar com dados nesse nível de detalhe, colocam-se, de imediato, dois problemas principais. O primeiro está no sigilo dos prontuários, que pode ser contornado por meio da disponibilização de dados anonimizados e estruturados de forma a atender às demandas de investigações clínicas ou licenciamento especificamente definido. Outro problema, mais técnico e de solução mais difícil, reside no emprego, por grande parte das Unidades Hospitalares (UH), de softwares para prontuários eletrônicos sem maior estruturação para a entrada dos dados, com muitos campos de texto livre, que dificultam análises e extração posteriores.

Somam-se a esses problemas, os desafios do desenvolvimento e implantação de uma infraestrutura que permita apoiar a publicação de dados FAIR. Embora a proposta dos princípios já tenha alguns anos, fornecendo uma base conceitual e diretrizes que se popularizaram rapidamente, ainda são poucas as alternativas tec-

nológicas já experimentadas em conjunto. Certamente, a utilização de abordagens e padrões da Web Semântica constitui um sólido aporte às soluções sendo prospectadas e desenvolvidas, mas mecanismos e tecnologias complementares ainda se fazem necessários. As próximas duas subseções descrevem a rede VODAN em mais detalhe, assim como alguns dos principais recursos e soluções tecnológicas que a apoiam.

2.1 A Rede de Implementação VODAN

A rede de implementação VODAN surgiu no início de 2020, como um esforço conjunto para implementar, experimentar e agilizar soluções (algumas já sendo experimentadas de forma independente em outros domínios) de apoio à publicação e exploração de dados FAIR no contexto das pesquisas associadas à COVID-19 e a outros surtos virais futuros. A rede propõe um esforço para a assim chamada *FAIRificação*[15] de dados COVID-19, mesmo pós fato, empregando o modelo FPC-OMS para estabelecer a padronização de informações (SATTI *et al.*, 2020). O processo de *FAIRificação* dos dados promove a aplicação dos princípios FAIR aos dados e metadados, assim como à infraestrutura de suporte a eles. Para isso, de modo geral, o processo contempla as etapas de: (i) coleta de dados não FAIR; (ii) análise dos dados coletados; (iii) definição de um modelo semântico para o conjunto de dados que permita descrever o significado de entidades e suas relações, com precisão e sem ambiguidade; (iv) definição de metadados associados aos dados coletados, incluindo, dentre outros, proveniência, distribuição e localização, tipo de acesso; (v) tratamento para tornar o dado potencialmente interligável com outras fontes, com atribuição de identificadores persistentes, anotação com base em vocabulários controlados e/ou ontologias, empregando tecnologias e padrões de Web Semântica e Dados Conectados (*Linked Data*) (HEATH; BIZER, 2011); (vi) definição dos metadados associados aos dados (e seu tratamento para que também sejam FAIR); publicação dos dados e seus metadados.

Após o processo de *FAIRificação*, obtém-se um conjunto de dados e metadados aderentes aos princípios FAIR. Esses dados e metadados bem estruturados poderão ser explorados através de mecanismos que utilizem técnicas de aprendizado de máquina e outras abordagens de inteligência artificial (IA) para descobrir padrões significativos nos surtos epidêmicos, apoiando decisões e ações para o seu enfrentamento. Como apresentado em (SATTI *et al.*, 2020), é vital garantir que os dados, metadados e vocabulários empregados sejam FAIR, no sentido original do acrôni-

15 FAIRificação dos dados – processo para a transformação de dados não FAIR em dados FAIR. Disponível em https://www.go-fair.org/fair-principles/fairification-process/

mo, mas também no sentido de "*Federated, AI- Ready*", ou seja, dados federados e prontos para IA.

Ao final dos desenvolvimentos associados à iniciativa VODAN, espera-se o estabelecimento de uma rede federada de FAIR DPs epidemiológicos, promovendo serviços e dados FAIR, acessíveis aos pesquisadores, para estudos sobre a pandemia de COVID-19 e outras epidemias que possam surgir no futuro.

A rede VODAN Africa&Asia[16] foi a primeira iniciativa de implementação da rede VODAN. É financiada pela Fundação Philips[17] e visa promover o acesso distribuído aos dados de FPC de países da África e da Ásia, para apoiar o combate à pandemia da COVID-19, atendendo às Universidades e aos Hospitais da África em Uganda, Etiópia, Nigéria, Quênia, Tunísia e Zimbábue, dentre outros países. Essa iniciativa direcionou suas atividades para o treinamento de pesquisadores e projetistas de dados na criação de FAIR DPs. Os treinamentos orientaram os participantes sobre os princípios FAIR e sobre o processo de construção dos FAIR DPs, garantindo que os dados e metadados publicados sejam interligados e possam estar disponíveis e serem processados por agentes de software. Como consequência, em 22 julho de 2020[18], foi disponibilizado o primeiro FAIR DP do mundo, em Uganda. Desde então, outros FAIR DPs vêm sendo ativados, a partir da *FAIRificação* dos dados e metadados provenientes das UHs parceiras.

Diferentemente da rede VODAN Africa&Asia, o projeto VODAN BR optou por um escopo menor, visando o desenvolvimento de um ambiente de apoio inicialmente voltado para dados de dois hospitais parceiros, adequando-os ao FPC-OMS e criando a infraestrutura computacional para sua divulgação através de um primeiro FAIR DP no Brasil. Posteriormente, outros hospitais serão contemplados, podendo já se valer dos resultados da experiência piloto conduzida e da infraestrutura desenvolvida e testada.

2.2 Web Semântica e os Princípios FAIR

A Web Semântica propõe que os dados na Web sejam definidos e conectados de forma a serem interpretados tanto por seres humanos quanto por máquinas, promovendo seu compartilhamento e reuso por aplicações, empresas e pela comunidade. Para atingir esse objetivo, a proposta de representação de dados conectados estabelece um conjunto de padrões e melhores práticas para publicação e interliga-

16 https://www.vodan-totafrica.info/

17 http://www.digitaljournal.com/pr/4626217

18 https://kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at-kampala-international-university_1595432235

ção de dados estruturados na Web, apoiando-se na anotação de dados em vocabulários controlados e ontologias, facilitando a idenficação de novas conexões entre itens de diferentes fontes de dados, visando formar um espaço de dados global, a assim chamada Web de Dados (HEATH; BIZER, 2011).

Os princípios FAIR, inicialmente voltados para a gestão de dados de pesquisa, vêm se somar a muito do que já é proposto para a Web Semântica, com o objetivo de tornar os objetos digitais localizáveis, acessíveis, interoperáveis e reutilizáveis. Em sua essência, esses princípios agregam aos padrões estabelecidos pela W3C[19] a importância do uso de metadados para facilitar a descoberta e o entendimento dos dados, principalmente por máquinas (agentes de software). Ressalta-se que os princípios FAIR não estabelecem padrões ou tecnologias de suporte, mas sim orientam a criação de dados e metadados FAIR.

Padrões de metadados e anotações de conteúdo são estabelecidos para promover o entendimento comum sobre o significado dos dados, garantindo a interpretação correta e o seu emprego de forma adequada. Para que esses metadados possam ser interpretados por máquinas, eles precisam estar localizáveis e estruturados. Os metadados acionáveis por máquinas, essenciais aos princípios FAIR, fizeram com que membros do GO FAIR e da RDA, a partir de 2018, fomentassem a discussão sobre o *Metadata for Machine* (M4M), em uma série de eventos[20] para avaliar o estado da arte e estimular a criação e reutilização de componentes de metadados e de *templates* de metadados para processamento por máquinas. Na implementação VODAN, o M4M tem atuado na padronização dos metadados referentes aos catálogos e aos conjuntos de dados (*datasets*) que serão disponibilizados via FAIR DPs, assim como em uma série de serviços associados a eles.

De todo modo, não é trivial explicitar inequivocamente uma semântica compartilhada sobre esses ativos digitais e nisto as ontologias têm um papel fundamental. Atualmente, as ontologias são consideradas em diversas áreas da computação (STUDER; BENJAMINS; FENSER 1998), sendo que duas dessas grandes áreas são: (i) na área de modelagem conceitual, onde, através do processo de análise ontológica, constroem-se modelos bem fundamentados em ontologias de topo; e (ii) na área de Web Semântica, onde se utilizam tanto ontologias leves, na linha de vocabulários, taxonomias e tesauros, quanto ontologias robustas, preferencialmente seguindo modelos bem fundamentados e representadas em linguagens expressivas, que possam ser exploradas por mecanismos de inferência, para geração de mais conhecimento.

19 W3C Semantic Web Activity https://www.w3.org/2013/data/

20 https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/

Considerando a definição de ontologia como "*... uma especificação formal e explícita de uma conceitualização compartilhada*" (SANTOS, 2020), depreende-se que: uma ontologia é considerada formal por ser interpretável por máquina; é explícita pois apresenta as especificações de conceitos, propriedades, relações, funções, restrições e axiomas muito bem definidas; é uma conceitualização por definir um modelo abstrato e uma visão sobre um fenômeno do mundo que se quer representar; e é compartilhada por ser um conhecimento consensual entre aqueles que trabalham com o domínio ou aplicações em questão.

A abordagem para assegurar o formalismo e a flexibilidade para a criação e disponibilização dos dados e metadados utiliza a linguagem RDF (*Resource Description Framework*) [21], incluindo nesse contexto o RDFS (*RDF Schema*) [22]. A linguagem OWL (*Web Ontology Language*) [23], desenvolvida para a criação de ontologias robustas, emprega também esse padrão.

O formalismo da especificação RDF está associado ao padrão estrutural utilizado para descrever e armazenar os dados. Esse padrão é definido por triplas constituídas dos seguintes elementos: <*sujeito*> <*predicado*> <*objeto*>. Cada tripla constitui uma declaração, unidade básica do RDF, é um conjunto de declarações que descreve um recurso da Web. Cada recurso, por sua vez, possui um identificador único denominado URI (*Universal Resource Identifier*). As URLs (*Uniform Resource Locators*) associadas às URIs são derreferenciadas, ou seja, podem ser acessadas por meio de navegadores, disponibilizando informações sobre o recurso. Esse identificador único permite o reuso de recursos entre diferentes fontes de dados, agilizando implementações, propiciando interoperabilidade e facilitando integrações.

Por descrever dados e seus metadados, o RDF permite uma flexibilidade na construção e na evolução de esquemas não disponíveis nos Sistemas Gerenciadores de Banco de Dados (SGBD) usualmente utilizados, a exemplo dos baseados tecnologias relacionais. O conjunto de declarações representadas por triplas RDF constituem um Grafo de Conhecimento RDF.

3. O projeto VODAN BR e a perspectiva da gestão de dados e metadados FAIR

O projeto VODAN BR estabeleceu um conjunto de premissas a serem respeitadas durante as suas fases de implementação. Essas premissas orientam as atividades relacionadas com a gestão de dados e seus metadados, visando estabelecer uma

21 https://www.w3.org/wiki/RDF
22 https://www.w3.org/TR/rdf-schema/
23 https://www.w3.org/OWL/

estrutura capaz de ser rapidamente ajustada, que reduz, significativamente, a necessidade de mudanças em aplicações/ferramentas a cada evolução e versionamento do FPC ou dos instrumentos terminológicos de referência. Dentre as premissas estabelecidas merecem destaque:

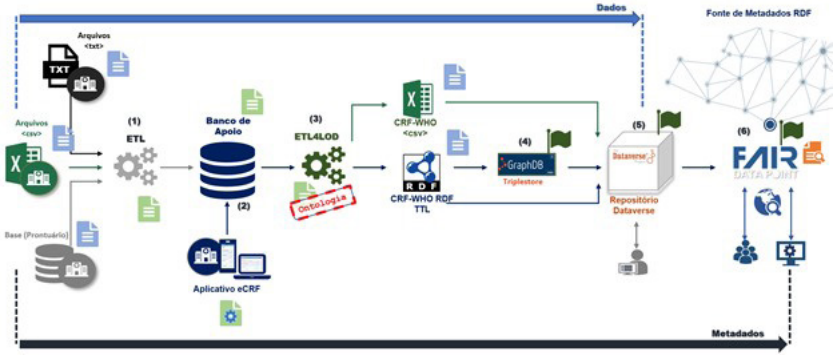
a) criar uma infraestrutura capaz de implementar e disponibilizar um FPC digital (aplicativo), centrado nos usuários dos serviços de saúde, que seja capaz de atender aos episódios epidêmicos dessa pandemia;
b) armazenar as informações estabelecidas no FPC-OMS, de forma anonimizada, considerando possíveis versões do FPC atual para inclusão, alteração ou exclusão de elementos do formulário;
c) possibilitar a criação de FPCs nacionais ou a inclusão de questões adicionais específicas. Essa necessidade foi apresentada, tendo em vista os diferentes tipos de formulários de pesquisas empregados no Brasil, que, além dos elementos estabelecidos pelo FPC-OMS, se preocupam com informações específicas, relevantes para pesquisas no País, como, por exemplo, participação em campanhas de vacinação e data da última dose;
d) promover uma modelagem conceitual que permita o alinhamento dos elementos dos formulários às ontologias (modelos semânticos), auxiliando o processo de *FAIRificação* dos dados;
e) prover uma infraestrutura flexível, modular, escalar e ágil, a fim de dar suporte a adaptações em softwares e bases de dados;
f) transformar os dados coletados, isto é, "dados não FAIR" em dados interligados, mapeando-os para formatos legíveis por máquina, utilizando RDF, disponibilizando-os em datasets e publicando seus metadados, também em RDF, em um FAIR DP;
g) disponibilizar publicamente um FAIR DP configurado para atender as condições acordadas com os participantes, possibilitando acesso aos dados por meio de consultas controladas e não por tradicionais downloads.

Respeitando essas premissas, foi concebida uma plataforma para tratamento dos dados e serviços que abrange, desde a disponibilização dos dados de pesquisa clínica pelas UHs até a publicação de metadados no FAIR DP. A plataforma, representada na Figura 1, tem como principais requisitos ser modular, distribuída, escalável e flexível. Modular, pois, as atividades previstas estão organizadas na forma de módulos que interagem de modo encadeado, sendo o resultado de um módulo a entrada do módulo subsequente. Escalável e distribuída pois a ideia é que um banco de dados de apoio seja disponibilizado em cada UH, assim como os repositórios

e/ou bancos do tipo *triplestore* hospedarão os dados estruturados de acordo com o FPC-OMS em suas diferentes distribuições ou formatos. Dessa forma, a medida que mais hospitais passem a participar do projeto, mais infraestrutura computacional será adicionada, fazendo com que ocorra um natural escalonamento horizontal. Além disso, é uma plataforma flexível, pois os dados heterogêneos produzidos pelas UHs são tratados e transformados para uma representação em grafo RDF, que é um dos formatos que facilita a interligação de dados.

Inicialmente, conforme representação na Figura 1, (1) a plataforma capta dados que podem estar em diversos formatos, como por exemplo, txt, csv, ou ainda, no formato utilizado em cada UH, e, via um processo de Extração-Transformação-Carga (ou ETL do inglês para *Extraction-Tranformation-Loading*) realiza a depuração e transformação dos dados, armazenando-os em um banco de dados de apoio (2), que pode também receber diretamente dados através de aplicativo móvel (eCRF) especificamente desenvolvido. Os dados armazenados no banco de dados de apoio passam então por uma transformação para dados conectados (3), sendo anotados em vocabulários e ontologias, para atender ao princípio de interoperabilidade. São então carregados em um banco de dados em grafo (4), no papel de um *triplestore*, ou, na forma de um *dataset* RDF disponibilizado para download em um repositório (5). Os metadados associados passam também por processo de tratamento e transformação (3) sendo carregados e disponibilizados em um FAIR DP (6).

Figura 1 - Representação da Plataforma VODAN BR



Fonte: Elaborado pelos autores.

Como estabelecido na rede VODAN, os *datasets* deverão ser "visitados" por algoritmos, respeitando o acesso estabelecido pelas UHs. Os metadados associados, contemplando, por exemplo, informações sobre a origem dos dados existentes, tipos de distribuição e a política de acesso, estarão disponibilizados e acessíveis no FAIR DP.

Dos elementos que compõem a plataforma, distinguem-se, pela relevância no projeto e pela atenção e desafios no tratamento dos dados: (i) os mecanismos para captação de dados, contemplando diferentes requisitos e sistemas das UHs; (ii) o banco de dados de apoio, responsável pelo armazenamento dos dados oriundos dessas fontes de dados heterogêneas; (iii) a ferramenta de suporte ao tratamento, transformação e anotação para dados e metadados interligados; (iv) as alternativas para publicação dos dados; e (v) a criação e alimentação do FAIR DP, parte da federação de pontos de acesso geral do VODAN internacional.

As tarefas desempenhadas e as escolhas tecnológicas para estes 5 elementos principais da plataforma VODAN BR são descritas nas subseções a seguir.

3.1. Captação de Dados

O projeto prevê três formas diferentes de captação de dados de pesquisas clínicas de pacientes com COVID-19:

1) pelo emprego de aplicativo (eCRF) criado para o registro das informações, de acordo com o FPC-OMS;
2) por meio de uma ferramenta ETL para carga e tratamento de dados anonimizados a partir de arquivos nos formatos txt ou csv disponibilizados pelas UHs;
3) por meio de processos de ETL conectando banco a banco, com o propósito de transferir as informações do prontuário para o banco de dados de apoio, no formato estabelecido pelo FPC-OMS.

A captação a partir de prontuários digitais já existentes representa um desafio adicional. Apesar do emprego do banco de dados de apoio pelas UHs, como banco de transição para o formato do FPC-OMS, e de todas as facilidades que ele oferece, um dos principais problemas na análise e extração de dados clínicos para pesquisa decorre da flexibilidade dos sistemas de prontuários existentes que habilitam campos textuais para o registro de determinados aspectos do tratamento. A falta de padronização neste registro e o grande volume desses dados não estruturados (contemplam cada procedimento feito no paciente, incluindo medicações e exames de laboratório), dificultam o processo de coleta e transformação, requerendo o apoio de um profissional de saúde para sua interpretação e recodificação. Esse problema não é novidade e tem sido uma constante em estudos de interoperabilidade de dados de tratamento de saúde (SANTOS, 2020; CRUZ *et al.,* 2009).

Outro aspecto importante considerado foi a diversidade de informações de proveniência de dados (CRUZ *et al.,* 2009) a serem gerenciadas.

3.2 Criação e Manutenção do(s) Banco(s) de Dados de Apoio

Em virtude da heterogeneidade das fontes dos dados e dos dados per se, optou-se pelo desenvolvimento de um banco de dados de apoio para o tratamento e formatação dos dados, visando adequá-los à estrutura do FPC-OMS e dar suporte e agilizar o processo de transformação para dados conectados.

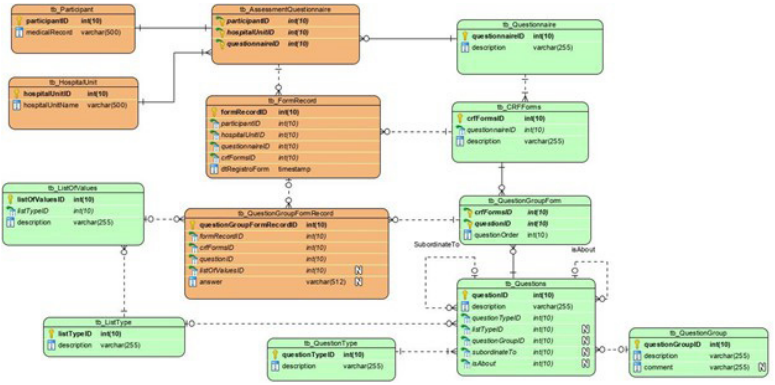
Destacamos que, apesar de parte das fontes de dados serem provenientes de prontuários de pacientes de sistemas até certo ponto semelhantes, optou-se por seguir uma modelagem aderente ao questionário para coleta de dados do FPC-OMS. Essa decisão foi crítica, pois permitiu o estabelecimento de uma estrutura de perguntas e respostas associada aos formulários/módulos orientada ao atendimento e, consequentemente, à coleta de informações. O formulário representa uma pesquisa, no nosso caso uma pesquisa de dados clínicos. Ele é constituído de um conjunto de questões, agrupadas em categorias bem definidas, coletadas por um agente de saúde, nesse caso em uma UH, considerando observações realizadas sobre um elemento de interesse, o paciente. Essa pesquisa requer uma visão espaço-temporal, tendo para isso: um *módulo de admissão*, destinado às questões do momento do acolhimento do paciente; um *módulo de acompanhamento*, destinado às questões sobre o paciente durante a internação; e um *módulo de desfecho*, com uma visão geral sobre o tratamento fornecido e a situação final do paciente.

O emprego de questões com respostas, em sua maioria, padronizadas, auxilia a descoberta de vocabulários e a adoção de tecnologias da Web Semântica, como por exemplo, ontologias que possam ser empregadas para definir um modelo semântico.

Outro aspecto importante refere-se à combinação da estrutura hierárquica Módulo/Agrupamento/Questão e Questão Subordinada que embute uma organização de conhecimento por categorias, permitindo definir visões diferenciadas para análise. Um exemplo de análise possível seria a avaliação da evolução dos casos (da admissão ao desfecho), considerando as comorbidades identificadas no momento de admissão e os medicamentos administrados durante a internação. O resultado dessa análise poderia auxiliar no processo de orientação de medicamentos indicados ou não em um tratamento, em face da comorbidade do paciente.

Para modelagem e implementação deste banco de dados de apoio, optou-se por utilizar uma modelagem baseada em tecnologia relacional, pela facilidade de manutenção e interação com os mecanismos e aplicativos de carga e manipulação de dados. Uma visão parcial do esquema deste banco de dados é apresentada na Figura 2, onde as entidades na cor verde representam a estrutura hierárquica do FPC-OMS e aquelas em laranja representam o registro das informações dos pacientes coletadas pelas unidades hospitalares.

Figura 2 - Extrato do Modelo Lógico do banco de dados de apoio – Visão FPC-OMS



Fonte: Elaborado pelos autores.

Nesse modelo, as informações sobre o FPC-OMS e suas versões são registradas na tabela *tb_Questionnaire*; os três módulos associados ao FPC são registrados na tabela *tb_CRFForms*; as questões são armazenadas na tabela *tb_Questions*, onde são associados o tipo da questão (*tb_QuestionType*), o agrupamento a que pertence (*tb_QuestionGroup*), se houver, e o tipo de lista de respostas padronizadas (*tb_ListType*), caso a questão exija uma resposta padronizada.

As informações dos participantes da pesquisa (pacientes) são lançadas na tabela *tb_Participant* e as da Unidade Hospitalar na tabela *tb_HospitalUnit*. A tabela *tb_AssessmentQuestionnaire* registra a abertura de um CRF-OMS para um participante. Para cada registro desta tabela, são lançados, na tabela *tb_FormRecord*, os módulos habilitados para o paciente. A tabela *tb_QuestionGroupFormRecord* armazena, por módulo, as questões e respostas obtidas da avaliação de cada paciente, considerando respostas textuais e respostas padronizadas (*tb_ListOfValues*).

3.3. Transformação para Dados Conectados

Após o tratamento e a carga dos dados de cada fonte (UH) em seu banco de dados de apoio, o próximo passo refere-se a um processo de transformação para um modelo semântico, seguindo o paradigma de dados conectados.

Para o projeto VODAN, a equipe do FAIR Data Team disponibilizou um modelo de dados semântico que representa a versão rápida do FPC-OMS. Esse modelo semântico (ou ontologia) foi denominado WHO-COVID-CRF[24] e, além da representação do FPC-OMS, estabeleceu um conjunto de entidades do domínio da área da saúde, às quais as questões do questionário estão relacionadas. Essas entidades

24 WHO-COVID-CRF: https://github.com/FAIRDataTeam/WHO-COVID-CRF

remetem a outras ontologias existentes e bem documentadas, proporcionando qualidade e informações adicionais para nortear os usuários no preenchimento da pesquisa.

Por ter sido desenvolvido orientado ao formulário FPC-OMS, a análise desse modelo semântico identificou uma série de similaridades que permitiram estender a modelagem do banco de dados de apoio, incluindo as informações da ontologia referentes à identificação, estruturação e valoração, de modo a agilizar o processo de *FAIRificação* dos dados.

A estruturação da ontologia WHO-COVID-CRF permitiu o emprego de suas informações para a carga inicial das tabelas que representam o questionário, com pouquíssimos ajustes. Por meio dessa carga, efetivou-se o alinhamento das informações das tabelas referentes ao questionário com a ontologia, permitindo a criação, pelo administrador de dados, de visões que apresentam o questionário e suas informações ontológicas, bem como, de visões que auxiliem a etapa de transformação para dados interligados realizada posteriormente.

Para transformação para dados conectados, optou-se pela ferramenta ETL-4LOD[25]. Essa ferramenta foi desenvolvida, inicialmente, por meio de uma parceria entre as universidades UFRJ e a UFES, no Projeto *LinkedDataBR*[26], tendo como meta a construção de uma infraestrutura de suporte à publicação de dados abertos empregando os padrões e tecnologias da Web Semântica. A ETL4LOD consiste em um conjunto de plugins, desenvolvidos em JAVA, que estendem as funcionalidades do *Pentaho Data Integration*, uma ferramenta ETL largamente usada, propiciando a transformação de dados de diferentes fontes para dados conectados.

Da mesma forma como a ferramenta vem sendo adaptada para o tratamento dos dados, também contempla o tratamento dos metadados, de forma a apoiar o processo de *FAIRificação* como um todo.

Convém ressaltar que a modelagem adotada permite que ontologias de interesse que venham a surgir sejam incorporadas ao banco, servindo para a realização de anotações adicionais que contribuirão com a redução de ambiguidade dos dados e metadados tratados.

3.4. Publicação dos Dados

Seguindo as orientações da rede VODAN, os dados das pesquisas devem ser disponibilizados no formato de dados conectados, usando o padrão RDF. Seguindo as tendências para gestão de dados de pesquisa e sua disponibilização em repositórios

25 ETL4LOD: disponível em https://github.com/johncurcio/ETL4LODPlus

26 https://memoria.rnp.br/_arquivo/gt/2010/GT-LinkedDataBR_fase1.pdf

institucionais ou temáticos, uma de nossas alternativas para publicação dos dados foi o uso de uma plataforma de repositório. No VODAN BR, optou-se pelo uso do *Dataverse*[27], por ser a plataforma previamente selecionada pela instituição coordenadora do GO FAIR Saúde Brasil, a Fundação Oswaldo Cruz, para a publicação dos dados de suas pesquisas.

O *Dataverse* é um repositório de dados de código aberto, desenvolvido pelo Instituto de Ciências Sociais Quantitativas de Harvard (IQSS), para armazenar, compartilhar, publicar, citar, explorar e analisar dados de pesquisa. O repositório hospeda vários arquivos virtuais chamados *dataverses*. Cada *dataverse* contém conjuntos de datasets, e cada *dataset* contém metadados e arquivos de dados descritivos (incluindo documentação e código que acompanham os dados). Como método de organização, os *dataverses* também podem conter outros *dataverses*.

Visando a ampliar o reuso dos dados, além dos *datasets* no padrão RDF, o projeto estabeleceu outros dois formatos de distribuição: o primeiro, em um *triplestore* apoiado por um SGBD em grafo utilizando a ferramenta *GraphDB*[28], e o segundo, no formato .csv, para um uso mais tradicional dos dados.

O *GraphDB* é um SGBD para bancos de dados em estrutura de grafos, também utilizado como *triplestore* RDF, que fornece uma estrutura ágil para publicação e consumo de dados conectados. Esse consumo é realizado por meio da linguagem SPARQL[29] (*SPARQL Protocol and RDF Query Language*), uma linguagem para consultas semânticas com um protocolo para acesso aos dados em RDF. Deste modo, em uma proposta inicial, cada hospital participante poderá ter seus dados disponibilizados em diferentes formatos e plataformas de distribuição, de acordo com sua conveniência e com o licenciamento que definir.

3.5. Publicação no FAIR DP VODAN BR

Conforme mencionado anteriormente, um FAIR DP é uma infraestrutura de armazenamento e acessibilidade a dados que tem como objetivos: (i) permitir que os proprietários de dados exponham seus *datasets* em conformidade com os princípios FAIR; (ii) facilitar a descoberta das informações sobre o FAIR DP pelos consumidores de dados, em uma rede de FAIR DPs; (iii) estabelecer mecanismos que gerenciem o acesso dos consumidores, de acordo com as licenças e restrições impostas aos dados por seus gestores; (iv) fornecer, aos proprietários dos dados, indicadores de acesso sobre os (meta)dados disponibilizados; e (v) proporcionar a

27 The Dataverse Project – Disponível em: https://dataverse.org

28 GraphDB. Disponível em: http://graphdb.ontotext.com/

29 SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/

interação dos dados para humanos, por meio da interface gráfica do usuário (em inglês *Graphical User Interface* – GUI), e para agentes de software, utilizando interface de programação de aplicativos (em inglês *Application Programming Interface* – API) (SANTOS *et al.*, 2016).

Para promover uma padronização para os FAIR DPs do VODAN, os metadados de referência foram estabelecidos e estruturados no modelo RDF pela equipe do FAIR Data Team. Essa padronização define um conjunto de metadados ricos que descrevem informações como, por exemplo, coerência estrutural e interna dos dados, licenças e fontes de referência, condições de acesso, contexto e proveniência (SANTOS, 2020).

Outro aspecto importante considerado foi a diversidade de informações de proveniência de dados (CRUZ *et al.*, 2020) a serem coletadas, gerenciadas e disponibilizadas no FAIR DP. Resumidamente, a proveniência de dados tem uma função muito importante em projetos de cunho científico ou mesmo comerciais. Ela pode ser definida como uma documentação histórica de um artefato (objeto, dado ou *dataset*) gerado por um procedimento conduzido por agente (pessoa, processo ou sistema computacional). Ela possibilita que estudiosos compreendam e sejam capazes de avaliar com maior precisão a importância e o contexto de criação, aplicação ou reuso daquele artefato. Proveniência é um tipo de metadado que aumenta as garantias da qualidade e a veracidade dos dados ou *datasets*. Ela auxilia na gestão dos dados do projeto como também oferece suporte na reprodutibilidade e confiabilidade.

A proveniência de dados no projeto VODAN BR poderá ser útil para os pesquisadores e profissionais de saúde que buscam compreender os efeitos da pandemia. Por exemplo, as informações de proveniência podem ser incorporadas no nível de registro, atribuindo descritores como parte do processo de transformação de dados (por exemplo, se um diagnóstico foi inserido por um médico ou derivado de uma versão do formulário ou se o dado é oriundo do processo de ETL). Esses detalhes são importantes porque *datasets* que incluem registros de uma fonte e não de outras ou incluem registros de várias fontes que não se distinguem em bancos de dados de uso geral e acabam gerando perfis e análises muito diferentes.

Seguindo as orientações da rede de implementação VODAN e os tutoriais elaborados pelo VODAN Africa&Asia, o FAIR DP do VODAN BR publicará os metadados referentes aos repositórios e seus *datasets*, descrevendo, em detalhes, as fontes de dados e seus itens. Passará, assim, a integrar a federação de FAIR DPs VODAN, que visa facilitar a divulgação/publicação de metadados sobre os dados do COVID-19, promovendo o acesso a esses dados por agentes de software e por humanos (SANTOS *et al.*, 2016).

4. Conclusão

O desenvolvimento de um ativo computacional sob a forma de uma plataforma para disponibilização de dados de pesquisas referentes aos surtos virais no meio de uma pandemia é um grande desafio, tanto no que diz respeito aos aspectos computacionais quanto aos aspectos de saúde pública. Como apresentado ao longo deste capítulo, o projeto VODAN BR vem trabalhando continuamente para a implementação de sua plataforma, mantendo uma visão geral sobre os dados, desde o momento de sua captação até a disponibilização dos metadados associados aos repositórios e aos *datasets* em FAIR DPs.

A experiência no projeto reforça a importância do FAIR DP na infraestrutura, não só como elemento essencial para a federação de pontos de acesso e mecanismos de busca e reuso a dados FAIR, mas também para apoio a dados sensíveis de pesquisa que requerem algum grau de sigilo, como é o caso dos dados sobre pacientes. Nesse aspecto, por meio dos FAIR DPs, o acesso aos metadados referentes aos dados de pesquisa são disponibilizados, dando-lhes visibilidade e acessibilidade. Entretanto, o acesso efetivo aos dados respeita condições bem definidas, promovendo "dados tão abertos quanto possíveis e tão fechados quanto necessários" (WILKINSON et al., 2016).

Entre as lições aprendidas quanto à plataforma estabelecida para o projeto VODAN BR, foi observada a carência de ferramentas que auxiliem o processo de *FAIRificação* como um todo. Algumas das soluções adotadas podem ser automatizadas, aprimorando o processo e tornando a plataforma mais estável para atender a novos desafios. Um exemplo passível de automatização ocorre no processo de publicação das distribuições de um *dataset* no repositório *Dataverse* e dos respectivos metadados associados no FAIR DP.

Por fim, temos objetivos finais semelhantes aos da rede VODAN Africa&Asia. Os desafios sendo vivenciados durante todas as fases do projeto propiciam visões diferentes e complementares, fornecendo uma riqueza de experiências que devem ser observadas e analisadas, para o estabelecimento de boas práticas a serem levadas para outras redes de implementação FAIR.

5. Agradecimentos

Este trabalho vem sendo elaborado através de múltiplos esforços. Os autores agradecem à equipe VODAN-BR, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, ao CNPq - Código de Financiamento 315399/2018-0, a FAPERJ, ao Hospital Federal Gaffreé Guinle, do Rio de Janeiro, ao Hospital Municipal São José, de Duque de Caxias ao professor Mauro Martin da ESDI-UERJ, e, em especial, ao grupo de alunos de graduação e

pós-graduação do programa PPGI/UFRJ e demais alunos voluntários que vêm se dedicando ao projeto desde abril de 2020.

6. Referências

DATA TOGETHER COVID-19 Appeal and Actions. Disponível em: https://www.go-fair.org/wp-content/uploads/2020/03/Data-Together-COVID-19-Statement-FINAL.pdf. Acesso em: 1 dez. 2020.

MANIFESTO VODAN IN. Disponível em: https://www.go-fair.org/wp-content/uploads/2020/03/VODAN-IN-Manifesto.pdf. Acesso em: 1 dez. 2020.

WILKINSON M et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**. v. 3 n. 160018. Disponível em: https://doi.org/10.1038/sdata.2016.18.2. Acesso em: 1 dez. 2020.

MONS, B. The VODAN IN: support of a FAIR-based infrastructure for COVID-19. **European Journal of Human Genetics**. v. 28. p.1-4. Disponível em: 10.1038/s41431-020-0635-7. Acesso em: 1 dez. 2020.

SATTI, F. et al. Semantic Bridge for Resolving Healthcare Data Interoperability *In:* INTERNATIONAL CONFERENCE ON INFORMATION NETWORKING, 2020, Barcelona, **Anais**..., 2020 p. 86-91.

HEATH, T.; BIZER, C. Linked Data: Evolving the Web into a Global Data Space. **Synthesis Lectures on the Semantic Web: Theory and Technology**, Morgan & Claypool, 2011. p. 1-136.

STUDER, R., BENJAMINS, R., FENSEL, D. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, n.1–2, p. 161–198, 2018.

SANTOS, L.O.B.S., FAIR DP Specification. Acesso em dezembro 2020. Disponível em: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec. Acesso em: 1 dez. 2020.

SANTOS, L.O.B.S., et al. **FAIR Data Points Supporting Big Data Interoperability, Enterprise Interoperability in the Digitized and Networked Factory of the Future**, Publisher: ISTE Press, 2016.

CRUZ, S. M.S, CAMPOS, M. L. M., MATTOSO. M. Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. 2009. *In:* International Conference on Web Services Los Angeles. **Anais...**United States: 2009. Disponível em: 10.1109/SERVICES-I.2009.18. Acesso em: 1 dez. 2020.

Como citar com o DOI individual

CAMPOS, Maria Luiza Machado; BORGES, Vania; Lopes, Giseli Rabello; Cavalcanti, Maria Claudia; MOREIRA, João; CRUZ, Sergio Manuel Serra da.VODAN BR – uma plataforma de apoio para dados COVID-19 seguindo os princípios FAIR. *In:* SALES, Luana Farias; VEIGA, Viviane dos Santos; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 253 - 270. DOI: 10.22477/9786589167242.cap18

Sobre os organizadores

Luana Farias Sales

Titulação mais alta e instituição por extenso e sigla: Doutora em Ciência da Informação pelo Programa de Pós-Graduação do IBICT/UFRJ (2011-2014). Mestre em Ciência da Informação pelo convênio UFF/IBICT (2004-2006), Graduação em Biblioteconomia e Documentação pela Universidade Federal Fluminense (2003).

INSTITUIÇÃO

Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT
Programa de Pós Graduação em Ciência da Informação - PPGCI
Programa de Pós-Graduação em Biblioteconomia da UNIRIO – PPGB

DADOS BIOGRÁFICOS:

Analista em C & T do MCTIC/IBICT, atuando como docente do Programa de Pós-graduação em Ciência da Informação do convênio IBICT-UFRJ e Coordenadora da Rede de Implementação do GO FAIR Brasil. Docente colaboradora no Programa de Pós-Graduação em Biblioteconomia (UNIRIO). Bolsista de Produtividade do CNPq Pq2. Líder do Grupo de Pesquisa BRIET: – Biblioteconomia, Recuperação, Interoperabilidade, E-science e Tecnologias.

E-mail: luanasales@ibict.br
CV: http://cnpq.br/9090064478702633
ORCID: http://orcid.org/0000-0002-3614-2356

Viviane Santos de Oliveira Veiga

Doutora em Ciências pelo Programa de Pós-Graduação em Informação e Comunicação em Saúde PPGICS/Fiocruz em (2017). Mestre em Saúde Pública pela Escola Nacional de Saúde pública Sérgio Arouca (2006). Bacharel em Biblioteconomia e Documentação pela Universidade Federal do Estado do Rio de Janeiro-UNIRIO (1999)

INSTITUIÇÃO

FUNDAÇÃO OSWALDO CRUZ

Rede de Bibliotecas Fiocruz
Programa de Pós-graduação em Informação e Comunicação em Saúde PPGICS/Fiocruz

DADOS BIOGRÁFICOS:

Pesquisadora na Fundação Oswaldo Cruz. Professora Permanente do Programa de Pós-graduação em Informação e Comunicação em Saúde PPGICS/Fiocruz. Professora convidada no Programa de Pós-Graduação em Saúde da Criança e da Mulher. Professora no curso de Especialização em Saúde do Trabalhador e Ecologia Humana Saúde do Trabalhador/ENSP/Fiocruz. Professora no curso de especialização em Informação Científica e Tecnológica em Saúde/ICICT/Fiocruz. Coordena a Rede de Bibliotecas Fiocruz e a Rede GO FAIR Brasil Saúde. Coordena e participa de grupos de pesquisa em Informação Científica e Tecnológica.
E-mail: viviane.veiga@icict.fiocruz.br
CV: http://lattes.cnpq.br/4983074089687751
ORCID: https://orcid.org/0000-0001-8318-7912

Patricia Corrêa Henning

Doutora em Ciências pelo Programa de Pós-Graduação em Informação e Comunicação em Saúde PPGICS/Fiocruz em (2013). Mestre em Ciência da Informação pelo convênio UFRJ/IBICT (1993). Graduada em Relações Internacionais pela Universidade de Brasília - UNB (1986)

INSTITUIÇÃO

Universidade Federal do Estado do Rio de Janeiro - UNIRIO
Programa de Pós-graduação em Saúde e Tecnologia no Espaço Hospitalar (PPGSTEH)

DADOS BIOGRÁFICOS:

Professora Associada aposentada do curso de graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (Unirio). Atualmente é professora visitante do Programa de Pós-graduação em Saúde e Tecnologia no Espaço Hospitalar (PPGSTEH) e uma das Coordenadoras da Rede GO FAIR Brasil Saúde - Enfermagem.
E-mail: patricia.henning@unirio.br
CV: http://lattes.cnpq.br/0970010723997242

ORCID: http://lattes.cnpq.br/0970010723997242

Luís Fernando Sayão

Doutor em Ciência da Informação pela ufrj/ibict (1994). Mestre em Ciência da Informação pela (ufrj/ibict). Graduação em Física pela Universidade Federal do Rio de Janeiro (1978).

INSTITUIÇÃO

Comissão Nacional de Energia Nuclear
Centro de Informação Nuclear
Programa de Pós-Graduação em Biblioteconomia da unirio - ppgb
Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa

DADOS BIOGRÁFICOS:

Tecnologista sênior desde 1980 na Comissão Nacional de Energia Nuclear. É conselheiro do conarq - Conselho Nacional de Arquivos, docente permanente do Programa de Pós-graduação em Ciência da Informação do convênio ibict-UFR. Docente Colaborador no Programa de Pós-Graduação em Biblioteconomia da unirio - Universidade Federal do Estado do Rio de Janeiro e no Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa. Bolsista de Produtividade do CNPq Pq2. Vice-líder do Grupo de Pesquisa briet – Biblioteconomia, Recuperação, Interoperabilidade, E-science e Tecnologias.

E-mail: luis.sayao@cnen.gov.br
CV: http://lattes.cnpq.br/342262312294838
ORCID: http://orcid.org/0000-0002-6970-0553

Comitê Editorial - Dados biográficos

Barend Mons

Universidade de Leiden (LUMC) - Holanda. Biólogo molecular por formação e um dos principais especialistas em dados do FAIR. Na primeira década de sua carreira científica, ele dedicou-se à pesquisa fundamental sobre os parasitas da malária e, posteriormente, à pesquisa translacional para vacinas contra a malária. No ano de 2000, passou a pesquisar sobre administração de dados e análise de sistemas biológicos. Professor em Leiden e mais conhecido por inovações em colaboração acadêmica, especialmente nanopublicações, descoberta baseada em gráficos de conhecimento e, mais recentemente, a iniciativa de dados FAIR e GO FAIR . Desde 2012, ele é professor de bio-semântica no Departamento de Genética Humana do Centro Médico da Universidade de Leiden (LUMC) na Holanda. Em 2015, Barend foi nomeado presidente do Grupo de alto nível de especialistas sobre a nuvem europeia de ciência aberta . Desde 2017, Barend dirige o escritório de Coordenação e Apoio Internacional da iniciativa GO FAIR. Ele também é o presidente eleito do CODATA , o comitê permanente de questões relacionadas a dados de pesquisa do Conselho Internacional de Ciência . Barend é membro da Academia Holandesa de Tecnologia e Inovação (ACTI). Representante europeu no Conselho de Dados e Informações de Pesquisa (BRDI) das Academias Nacionais de Ciência, Engenharia e Medicinanos Estados Unidos. Palestrante frequente sobre FAIR e ciência aberta em todo o mundo, e participa de vários conselhos consultivos científicos de projetos de pesquisa internacionais

Email: barendmons@gmail.com
ORCID: https://orcid.org/0000-0003-3934-0072

Luiz Olavo Bonino

Universidade de Twente - Holanda. Responsável pelo desenvolvimento de várias tecnologias e ferramentas de apoio a criação, publicação, indexação, pesquisa, avaliação e anotação de (meta) dados FAIR e professor associado do grupo BioSemantics no Centro Médico da Universidade de Leiden, na Holanda. Mestrado em Informática pela Universidade Federal do Espírito Santo (2004). Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Computação. Atuando principalmente nos seguintes temas:Comércio eletrônico, Arquitetura de

sistemas de informação, Business modeling, Agentes inteligentes. Mestrado em Informática, graduação em Turismo. Seus idiomas são: Inglês, Espanhol e Italiano.
Email: luiz.bonino@go-fair.org
ORCID: https://orcid.org/0000-0002-1164-1351

Giancarlo Guizzardi

Universidade de Twente - Holanda. Cientista da computação brasileiro-italiano especializado em modelagem conceitual, modelagem corporativa, ontologia aplicada e sistemas de informação orientados por ontologia. Possui doutorado (com a mais alta distinção) pela University of Twente (2005), Holanda e pós-doutorado pela University of Trento (2013-2015), Itália. Atualmente é professor adjunto da Universidade Federal do Espírito onde coordena o grupo NEMO, um grupo de pesquisas em Ontologias e Modelagem Conceitual com cerca de 50 membros. Desde 2003 tem sido cientista visitante, colaborador de Pesquisa e Pesquisador Associado ao Laboratório de Ontologia Aplicada (LOA)-Instituto de Ciências e Tecnologia da Cognição (ISTC) em Trento, Itália. É revisor de Periódicos como MISQ Quarterly, Information Systems, IEEE Transactions on Data and Knowledge Engineering, IEEE Transactions of Software Engineering, Journal of Software and System Modeling, Journal of Data Semantics entre outros. É Editor Associado do Journal of Applied Ontology e membro do corpo editorial do Requirements Engineering Journal e Enterprise Modelling and Information Systems Architectures e foi membro do corpo editorial do Semantic Web Journal (SWJ) (entre 2009-2014), entre outros. Entre 2005-2007 e novamente a partir de 2013 foi pesquisador associado do CNR (Consiglio Nazionalle delle Ricerche), Itália. Foi por duas vezes eleito como membro do Executive Council da International Association for Ontology and Applications (IAOA). É atualmente membro do Advisory Board da IAOA e do Advisory Board do Ontology Summit e é co-chair IAOA Special Interest Group (SIG) on Ontologies and Conceptual Modeling.(e.g.,BPM 2016,ER 2014, CIbSE 2015, ENMO 2015, BalticDB&IS'06, Webmidia/SBSC/LAWeb 2008, ODISE 2012, i*Star 2014, SoEEEA'10, ONTOBRAS 2012, ISKO Brasil 2011), Invited Speaker (SLE 2012, IAOA Ontology Summer School 2012, CONSEGI'2011, Conceptual Space at Work 2012, Conferência Web.Br 2010 e 2014), invited panelist (e.g., ER 2014, CAISE 2012, UNDP Global Meeting on Government Interoperability Networks 2010, Ontology Summit 2012, Ontolog Earth Series 2013). Em particular, 2012 foi convidado para apresentar seu trabalho de pesquisa em um Dagstuhl Seminar sobre Cognitive Approaches to the Semantic Web e, novamente, em 2013 em um Dagstuhl Seminar sobre Automated Reasoning in Conceptual Schemas e em 2014 em um Dagstuhl Seminar sobre Spatial Semantics and Robotics. Por fim, é consultor científico da

European Commission.
Email: gguizzardi@gmail.com
ORCID: https://orcid.org/0000-0002-3452-553X

Maria Luiza Machado Campos

Possui graduação em Engenharia Civil pela Universidade Federal do Rio Grande do Sul (1978), mestrado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (1984) e doutorado em Information Systems - University Of East Anglia (1993), Inglaterra. Pós-doutorado no Laboratory of Applied Ontology, CNRS, Itália (2015). Atualmente é professora no Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro, e uma das coordenadoras do grupo de pesquisas GRECO, atuando como pesquisadora e orientadora de mestrado e doutorado no Programa de Pós-graduação em Informática da mesma universidade. Foi coordenadora do Bacharelado em Ciência da Computação e do Programa de Pós-graduação em Informática, assim como Diretora Adjunta de Extensão do Instituto de Matemática da UFRJ. Seus principais temas de pesquisa estão associados à integração de informações heterogêneas, abordando principalmente os seguintes temas: metadados, ontologias, modelagem conceitual, banco de dados, data warehousing e web semântica. Dentre as áreas de aplicação em que tem trabalhado com mais frequência, incluem-se: bioinformática, e-gov, e-ciência e sistemas de emergência.
Email: mluiza.campos@gmail.com
CV: http://lattes.cnpq.br/0659658820912418
ORCID: 0000-0002-7930-612X

Abel Parcker

Diretor do Programa SciELO / FAPESP e coordenador de Projetos da FapUNIFESP. Packer é um dos fundadores da SciELO (Scientific Electronic Library Online, Biblioteca Científica Eletrônica Online), plataforma de acesso público a periódicos científicos que, neste ano, completa 20 anos. Lançada em 1998, a plataforma foi concebida como uma estratégia para superar o fenômeno conhecido como "ciência perdida", causado pela presença muito fraca dos periódicos de países em desenvolvimento nos índices internacionais.

Pioneiro no movimento internacional de acesso aberto a publicações científicas, ao longo desse período, a SciELO tornou-se parte essencial da infraestrutura da pesquisa na maioria dos países em que opera. Também é utilizado em muitos países como referência em avaliação de pesquisas, como um complemento das avaliações realizadas com base em índices internacionais - tendo se tornado, assim, um pa-

drão de qualidade. Atualmente, a plataforma disponibiliza 1285 periódicos e 745.182 artigos, e computa quase 17 milhões de citações.
Email: abel.packer@scielo.org
ORCID: https://orcid.org/0000-0001-9610-5728

Carlos Roberto Lyra da Silva

Graduado em Enfermagem e Obstetrícia pela Universidade Federal do Estado do Rio de Janeiro (1994), Mestre em Enfermagem pela Universidade Federal do Estado do Rio de Janeiro (2000), Doutor em Enfermagem pela Universidade Federal do Rio de Janeiro (2008) e Pós-Doutor pelo Programa Associado de Pós-Graduação em Enfermagem da UPE/UEPB. Diretor da Diretoria de Pós-Graduação da Pró-Reitoria de Pós-Graduação, Pesquisa e Inovação da Universidade Federal do Estado do Rio de Janeiro - PROPGPI/UNIRIO (atual). Tem experiência em Fundamentos de Enfermagem, atuando principalmente nos seguintes temas: enfermagem, cuidado de enfermagem, UTI, conforto e tecnologia. Concluiu o Curso preparatório para o Portal Web of Knowledge e as bases de dados promovido pelo IBICT-RJ. Editor Gerente da Revista de Pesquisa Cuidado é Fundamental Online. Professor Associado do Departamento de Enfermagem Fundamental da Escola de Enfermagem Alfredo Pinto - DEF/EEAP. Docente Permanente dos Programas de Pós-Graduação em Enfermagem - PPGENF e de Enfermagem e Biociências PPGENFBIO da UNIRIO. Orientador dos Programas de Mestrado e Doutorado em Enfermagem da EEAP/UNIRIO. Foi Membro do Comitê de Ética em Pesquisa e da Câmara de Pesquisa e de Bolsas da UNIRIO. Consultor ad hoc externo do Departamento de Pesquisa da UFU. Pertence ao Banco Nacional de Avaliadores de Programas de Residência Multiprofissional em Área Profissional da Saúde da CNRMS - Ministério da Educação. Foi Coordenador do Curso de Mestrado em Enfermagem do PPGENF-UNIRIO.
Email: profunirio@gmail.com
CV: http://lattes.cnpq.br/5699679119049526
ORCID: https://orcid.org/0000-0002-4327-6272

Teresa Tonini

Doutorado em Saúde Coletiva (2006) pelo Instituto de Medicina Social (IMS) da Universidade do Estado do Rio de Janeiro (UERJ). Mestrado em Enfermagem (1999) e Graduação em Enfermagem e Obstetrícia (1986) pela Escola de Enfermagem Anna Nery (EEAN) da Universidade Federal do Rio de Janeiro (UFRJ). Professora Associada do Departamento de Enfermagem Fundamental; Coordenadora do Programa de Pós-Graduação em Enfermagem e Biociências

(PPGENFBIO). Gestão 2013-2016; Coordenadora do Programa de Pós-Graduação em Enfermagem (PPGENF). Gestão 2010-2013 da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Representante Institucional na Red Internacional Centros Colaboradores de la Fundación Index. Membro da Comissão Acadêmica do Colégio Doutoral de Enfermagem do Grupo Tordesilhas. Coordenadora do convênio interinstitucional com a Universidad Sur Colombiana USCO (Neiva/Colômbia) e do convênio interinstitucional com a Universidade Estadual de Roraima (UERR). Assessora Internacional da USCO para Implantação do Curso de Doutorado em Ciências de La Salud. Conselheira Suplente do Conselho Regional de Enfermagem (COREN-RJ) - Gestão 2015-2017. Experiência nas áreas de Enfermagem e de Saúde Coletiva, com objetos de estudos sobre os seguintes temas: enfermagem, cuidados de enfermagem, gerência/administração em enfermagem, avaliação em saúde, avaliação dos serviços de enfermagem, segurança do paciente.
Email: teresa.tonini@unirio.br
CV: http://lattes.cnpq.br/3691852768131499
ORCID: https://orcid.org/0000-0002-5253-2485

Silvana Aparecida Borsetti Gregório Vidotti

Licenciada em Matemática pelo Instituto de Biociências, Letras e Ciências Exatas da UNESP (1986). Especialista em Ciência da Computação pelo Instituto de Ciências Matemáticas de São Carlos da USP (1987). Mestre em Ciências - área de concentração - Ciências da Computação e Matemática Computacional - pelo Instituto de Ciências Matemáticas de São Carlos da USP (1993). Doutora em Educação - área de concentração Educação Brasileira - pela Faculdade de Filosofia e Ciências da UNESP (2001). Atuação profissional: Professora Assistente-Doutora em Regime de Dedicação Integral à Docência e à Pesquisa da Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Filosofia e Ciências - FFC - Campus de Marília, Departamento de Ciência da Informação. Docente dos cursos de graduação em Arquivologia e Biblioteconomia e dos cursos de mestrado acadêmico e doutorado em Ciência da Informação da Unesp. Coordenadora do Programa de Pós-Graduação em Ciência da Informação da Unesp (2004 - 2011) Coordenadora do Doutorado Interinstitucional (DINTER) Unesp e Universidade Federal do Ceará (2010-2014) Assessora da Pró-Reitoria de Pós-Graduação da Unesp - PROPG (2013-2017). Assessora da Pró-Reitoria de Graduação da Unesp - PROGRAD (inicio: 2017). Parecerista ad hoc de agências de fomento nacionais e membro de Comitês Científicos de periódicos científicos Membro Titular do Conselho de Gestão Científica do Núcleo de Computação Científica da Unesp - GridUnesp. Coordenadora do Laboratório de Desenvolvimento e Aplicação de Multímidia da FFC

- UNESP. Coordenadora acadêmica do Repositório Institucional Uneso (início: 2014). Coordenadora do Portal Docente Unesp (início: 2018). Coordenadora do Comitê Gestor de Acesso Aberto da Unesp (início: 2019). Membro da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Líder do Grupo de Pesquisa - Novas Tecnologias em Informação (GP-NTI) Bolsista em Produtividade em Pesquisa - CNPq/PQ no período de 2014 a 2020. Pesquisadora da área de Ciência da Informação, com ênfases em Tecnologias de Informação e Comunicação e em Arquitetura da Informação digital, Acessibilidade, Usabilidade e Experiência de Usuário.
Email: svidotti@gmail.com
CV: http://lattes.cnpq.br/7390573927636069
ORCID: https://orcid.org/0000-0002-4216-0374

Fábio Gouveia

Tecnologista em Saúde Pública da Fundação Oswaldo Cruz - Brasil, líder do Grupo de Pesquisa Ciência, Dados, Redes e Metrias - (Scimetrics) e pesquisador na Rede Zika Ciências Sociais (https://fiocruz.tghn.org/zikanetwork/). Biólogo, mestre em Microbiologia e Imunologia e doutor em Química Biológica (Educação, Gestão e Difusão de Biociências), fez um pós-doutoramento curto como Visiting Fellow da Katolieke Universiteit Leuven (Bélgica) selecionado no edital 2009 do Coimbra Group Scholarships Programme for Young Professors and Researchers from Latin American Universities. Foi o ganhador, junto com Elaine Rabello, do Altmetric Research Award for Promising Altmetrics Research de 2020. É docente permanente do Programa de Pós-Graduação em Ciência da Informação do convênio IBICT/UFRJ e do Mestrado em Divulgação da Ciência, Tecnologia e Saúde da Fiocruz. Desenvolve pesquisas na área Ciência da Informação com ênfase em Estudos Métricos da Informação (Cientometria, Webometria, Altimetria e Indicadores de Ciência, Tecnologia e Inovação), Métodos Digitais, STS, Data Science e Tecnologia Blockchain, e na área de Divulgação Científica e Comunicação em Saúde, com ênfase em estudos sobre internet e mídias sociais.
Email: fgouveia@gmail.com
CV: http://lattes.cnpq.br/0733908324235348
ORCID: https://orcid.org/0000-0002-0082-2392

Gustavo Silva Saldanha

Pesquisador Titular do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Professor Adjunto da Universidade Federal do Estado do Rio de Janeiro (UNIRIO), bolsista de produtividade 2 do CNPq (2016-2018; 2019-2021),

bolsista Jovem Cientista do Estado da FAPERJ (2019-2021). Atua como docente nos programas de pós-graduação em Ciência da Informação do IBICT e em Biblioteconomia da UNIRIO. É líder, desde 2011, do grupo de pesquisa Ecce Liber: filosofia, linguagem e organização dos saberes (IBICT-UNIRIO). É editor executivo do periódico LIINC EM REVISTA. É membro, desde 2019, do Círculo Iberoamericano de Ciencia de la Información Documental (CIIBERCID); desde 2017, da equipe de pesquisadores Médiations en information communication spécialisée do Laboratoire dÉtudes et de Recherches Appliquées en Sciences Sociales (Lerass) da Université Toulouse III Paul Sabatier, França; desde 2008, da Rede Franco-Brasileira de Pesquisadores em Mediação e Usos Sociais dos Saberes e da Informação (Rede Mussi) e, desde 2014, do International Center for Information Ethics (ICIE). É co-fundador do Fórum Internacional A Arte da Bibliografia (2014) e do Fórum de Estudos Críticos da Informação (iKritika) (2013). Foi vice-coordenador na gestão 2015-2016 do Grupo de Trabalho (Estudos históricos e epistemológicos da Ciência da Informação) da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Atuou como bibliotecário da Fundação Biblioteca Nacional (2006-2010) (FBN) e do Instituto Brasileiro de Geografia e Estatística (2010-2012) (IBGE). Possui graduação em Biblioteconomia pela Universidade Federal de Minas Gerais (2006), especialização em Filosofia Medieval pela Faculdade São Bento-RJ (2010), mestrado em Ciência da Informação pela UFMG (2008), doutorado em Ciência da Informação pelo convênio IBICT-UFRJ (2012). Realizou, sob o fomento da Capes, no período 2017-2018, o estágio pós-doutoral na Université Toulouse III, Toulouse, França

Email: saldanhaquim@gmail.com
CV: http://lattes.cnpq.br/6143079905555041
ORCID: http://orcid.org/0000-0002-7679-8552

Realização

Cooperação

Financiamento

Esta obra é parte da Coleção PPGCI 50 Anos e foi composta em Minion pelo Programa de Educação tutorial da Escola de Comunicação da UFRJ em setembro de 2021.

“O presente livro é uma tentativa de reunir iniciativas brasileiras teóricas e empíricas em torno da aplicação dos princípios FAIR e ser mais um instrumento de disseminação desses princípios no Brasil, especialmente no âmbito da pesquisa em Ciência da Informação e da Computação.”

ISBN: 978-65-89167-24-2

EM COOPERAÇÃO

United Nations
Educational, Scientific and
Cultural Organization