DOI: 10.21680/2447-7842.2021v7n1ID25703

Mapeamento e descrição de características de repositórios multidisciplinares de dados científicos abertos

Mapping and description of characteristics of open scientific data multidisciplinary repositories

Elaine Rosangela de Oliveira Lucas [1]

Antonio Carlos Picalho [2]

Vitória Maria Hartmann Caitano [3]

Submetido em: 25-06-2021	Aceito em: 11-10-2021

Resumo: O compartilhamento de dados científicos possibilita acelerar a construção de novos conhecimentos, aumentando a eficiência dos recursos, ampliando as possibilidades de reuso e de reprodutibilidade de pesquisas. Com a finalidade de apresentar um conjunto de informações com características de repositórios de dados científicos abertos, definiu-se como objetivo para este estudo caracterizar e mapear os repositórios de dados científicos abertos como alternativa para a tomada de decisão institucional no campo da gestão de dados científicos. Realizou-se um estudo exploratório-descritivo, qualitativo e, com relação aos procedimentos técnicos, um levantamento bibliográfico-documental, em que foram analisados seis repositórios de dados científicos abertos: DANS, DataHub, Harvard-Dataverse, Dryad, Figshare e Zenodo. A partir de informações coletadas dos sites oficiais de cada um desses repositórios entre abril e maio de 2021, foram observadas características referentes às seguintes categorias de análise: fundação e responsabilidade pelo mantenimento do repositório; idiomas da plataforma e dos

[1] Doutora em Ciência da Informação - USP; Mestra em Engenharia de Produção - UFSC; Graduada em Biblioteconomia - UFSC.
[2] Mestrando em Engenharia e Gestão do Conhecimento - UFSC; Graduada em Biblioteconomia - UFSC.
[3] Graduanda em Biblioteconomia - UFSC.

datasets aceitos, mídias sociais e possibilidade de compartilhamento direto de conteúdo; tipos e formatos de arquivos aceitos, limites de uso, curadoria de conteúdo; identificadores persistentes de conteúdo, de autores e informações acerca de financiamento de pesquisa; controle de versões; métricas de uso; integração com outros aplicativos e custos. A análise dos sites dos repositórios gerou um quadro, disponibilizado no repositório Zenodo, com a descrição dos dados de cada repositório a partir das categorias. Com os resultados apresentados pretende-se auxiliar os pesquisadores na sua escolha por um repositório de dados científicos abertos que permita o compartilhamento dos dados da pesquisa.

Palavras-chave: Repositórios de dados científicos. Dados científicos abertos. Repositórios multidisciplinares.

1 INTRODUÇÃO

O aumento do volume de dados científicos tem sido expressivo nas últimas décadas e, como consequência, fez-se necessário o desenvolvimento de tecnologias e ferramentas para a organização e compartilhamento dessas informações, gerando assim o surgimento de termos como *e-Science*[4], repositórios digitais e/ou curadoria de dados científicos. Tais adventos surgiram com o intuito de acelerar a construção de novos conhecimentos e aumentar a eficiência dos investimentos das agências de fomento, por meio do compartilhamento de dados científicos entre pesquisadores.

O movimento da ciência aberta (*open science*), nascido na década de 1990, tem, atualmente, como um dos objetivos fornecer o livre acesso aos dados de pesquisa, respeitando os direitos autorais, períodos de embargo, questões de confidencialidade, privacidade, especificidades das áreas científicas, entre outros aspectos (CIUFFO *et al.*, 2017). “O conceito de ciência aberta representa uma nova

[4] O termo e-Science, originário no Reino Unido, foi cunhado por John Taylor, no ano de 1999. O termo adquiriu um significado que “representa a potência da ciência melhorada com o uso intensivo das TICs e sua ampliação em torno de um esforço colaborativo”. (FERREIRA, 2018, p. 15)

abordagem para o processo científico, com base no trabalho cooperativo e diferentes maneiras de difundir o conhecimento usando tecnologias digitais e novas ferramentas de colaboração" (EUROPEAN COMMISSION, 2016, p. 33, tradução nossa).

Isso permite a aceleração dos passos das pesquisas e um consequente avanço das descobertas científicas.

Entende-se como dados científicos "todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito pela comunidade científica como necessário para validar os resultados de pesquisa" (SILVA *et al.*, 2019, p. 308), o que não deve ser confundido com a produção acadêmico-científica tradicional (artigos, trabalhos, capítulos de livros etc.), nem com dados governamentais. Portanto, os termos 'dados abertos de pesquisa' ou 'dados científicos abertos' remetem ao uso, reuso e redistribuição desses dados por qualquer pessoa, desde que haja o respeito aos direitos autorais, com a devida citação dos autores. Conforme definição da Organização de Cooperação e de Desenvolvimento Econômico (OCDE) (2007, tradução nossa), dados de pesquisa são registros de fatos (pontuações numéricas, registros textuais, imagens e sons) usados como fontes primárias para a pesquisa científica, e que geralmente são aceitos na comunidade científica como necessários para validar os resultados da pesquisa.

Os dados e as coleções de dados de pesquisa duram mais que os projetos dos quais foram oriundos. Isso assegura que, mesmo findos os projetos, a posteriori outros pesquisadores podem se beneficiar e fazer uso dos dados produzidos. Assim, futuros projetos de pesquisa podem agregar novas descobertas ou elementos a esses dados, dando início a um novo ciclo (SAYÃO; SALES, 2016).

A importância do compartilhamento de dados abertos de pesquisa é crucial para uma ciência mais colaborativa, transparente e acessível. A colaboração evita a duplicação de esforços e, portanto, aumenta a eficiência. A transparência promove os resultados da investigação, que pode ser replicada e validada, reduzindo assim as chances de fraude científica, e, por fim, a acessibilidade facilita a participação dos

cidadãos e das empresas ao acelerar a inovação e divulgação de métodos ou resultados científicos que possam promover novos produtos e/ou serviços.

De acordo com Lecardelli (2020, p. 104), “compartilhar e disponibilizar dados científicos com qualidade é uma forma de cooperar com o bem estar social de forma ampla e globalizada”. A autora cita ainda a pandemia da Covid-19 como um exemplo relevante da ciência pela colaboração entre pesquisadores de todos os países. A corrida pelo desenvolvimento de vacinas poderia ter acontecido de forma ainda mais acelerada, caso a prática do compartilhamento de dados abertos de pesquisa fosse algo empregado por todos os pesquisadores.

Em outubro de 2020, Tedros Ghebreyesus Adhanom, diretor-geral da OMS (Organização Mundial da Saúde), junto com Michelle Bachelet, alta comissária da ONU (Organização das Nações Unidas), e Audrey Azoulay, diretora-geral da Unesco (Organização das Nações Unidas para a Educação, a Ciência e a Cultura), lançaram um apelo conjunto a favor da “ciência aberta”, qualificando-a de “questão fundamental de direitos humanos” que poderia garantir o direito ao acesso universal ao progresso científico e suas aplicações (ORGANIZAÇÃO DAS NAÇÕES UNIDAS, 2020, tradução nossa). Como resultado, muitas vacinas foram concluídas, acelerando o índice de produção e vacinação da população mundial e refletindo num possível fim da pandemia da COVID-19.

No contexto brasileiro, é relevante citar a iniciativa da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) junto a Universidade de São Paulo (USP), com apoio inicial do Instituto Fleury, Hospital Sírio-Libanês e Hospital Israelita Albert Einstein, que criaram em 2020 o repositório COVID-19 Data Sharing/BR. Pioneiro na América Latina, o repositório de dados abertos em um ano de atividade, reuniu mais de 50 milhões de dados de 800 mil pacientes (FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO, 2021). Com intuito de motivar o compartilhamento de dados pela comunidade científica, o repositório não só contribui para pesquisas sobre a doença como também fortalece a importância do fomento a cultura de compartilhamento de dados científicos abertos (COVID-19 DATA SHARING/BR, 2021).

A maioria das universidades e centros de pesquisa ao redor do mundo já tem repositórios institucionais que armazenam os resultados de pesquisa de seus membros (principalmente artigos, documentos de conferências, teses e dissertações). Além disso, em breve, tal como recomendado pela OECD (*Organization for Economic Cooperation and Development*), pela Comissão Europeia e-IRG (e-Infrastructure Reflection Group) e pela NSF (National Science Foundation), será essencial criar repositórios de dados de pesquisa que contêm o conjunto de dados (*datasets*) que originou cada investigação.

Assim sendo, a tendência de que pesquisadores compartilhem dados de suas pesquisas e desenvolvam planos para gestão e preservação desses dados propende a seguir aumentando. A capacidade tecnológica atual oferece a possibilidade de gerar grandes conjuntos de dados que, se foram bem abastecidos, bem estruturados e de fácil de acesso, permitem não só validar os resultados da pesquisa que geraram, mas também a possibilidade de que outros pesquisadores possam reutilizá-los para adicioná-los a outros conjuntos de dados ou para posterior análise, não previstos inicialmente pelos autores originais que compilaram os dados.

Com necessidade semelhante à do restante do mundo, o cenário brasileiro referente à ciência aberta, os centros de pesquisa e as universidades brasileiras também precisam lidar com essa grande demanda, pois são dessas instituições que grande parte do volume de dados advém, e a tendência ao longo dos próximos anos é que o ritmo se mantenha em escala crescente. Dito isso, as universidades brasileiras se veem diante de duas opções: construir o próprio repositório (institucional) ou indicar repositórios externos (temáticos), os quais seus pesquisadores possam utilizar para disponibilizar os dados das suas pesquisas.

De acordo com o estudo desenvolvido na Universidad Carlos III de Madrid (UC3M), indicar um repositório externo pode ser mais viável ao se levar em consideração dois fatores: o primeiro deles relacionado a todos os custos operacionais e questões administrativas ao iniciar e manter um repositório institucional de dados abertos de pesquisa e o segundo associado à variedade de formatos, tamanhos e características desses dados, o que os torna muito mais

representativos dentro de cada campo (tema), sendo agrupados por áreas de assunto ao invés de por instituições (ORTIZ-REPISO; HERNÁNDEZ-PEREZ, 2017).

Os resultados da pesquisa de Monteiro (2019) demonstraram que o cenário brasileiro sobre o tema não é apenas próximo nas necessidades como também "já possui importantes iniciativas e políticas que estimulam e, em alguns casos, exigem a preservação e o compartilhamento de dados científicos", mas que mesmo assim ainda "carece de continuidade e ampliação de estudos e práticas para melhoria tanto das estipulações quanto do suporte fornecido aos pesquisadores". Conhecer um repositório adequado e todas as suas possibilidades é um dos passos necessários para dar continuidade a todo esse processo de disponibilização dos dados abertos das pesquisas.

Visando entender quais são as semelhanças e diferenças nas funcionalidades dos repositórios multidisciplinares de dados científicos abertos disponíveis para uso pelas universidades brasileiras, este estudo tem por objetivo caracterizar e mapear os repositórios de dados científicos abertos como alternativa para a tomada de decisão institucional no campo da gestão de dados científicos. Os resultados da pesquisa permitirão que pesquisadores e universidades em busca de um repositório adequado à realidade da sua instituição tenham acesso a um quadro com informações de seis repositórios, facilitando, portanto, uma possível tomada de decisão ao definir as diretrizes de depósito de dados científicos abertos dos pesquisadores vinculados à instituição, criando padrões e evitando duplicidade de esforços.

2 REPOSITÓRIOS DE DADOS CIENTÍFICOS ABERTOS

Diante das novas perspectivas de trabalho com a incorporação das tecnologias de informação, além do interesse em coletivizar o conhecimento em forma de artigos, livros, anais, surge a possibilidade de se distribuir o conjunto de dados produzidos antes, durante e após o processo de uma pesquisa científica.

Tal ideia já se fazia presente em meados dos anos 1980, conforme Fienberg, Martin e Straf (1985), ao proporem caminhos para a distribuição de dados de pesquisa nos Estamos Unidos da América. Até então, muitas propostas esbarravam nas limitações técnicas, maiormente amparadas nos altos investimentos requeridos na época para implantação da estrutura necessária. Hoje, com a expansão e barateamento das tecnologias de informação, a discussão tem sido ampliada para a promoção de estratégias políticas, técnicas e operacionais que vislumbram uma estrutura universal de processamento de distribuição de dados científicos, conceito este presente na literatura sobre *e-science* e *data science*. Essa estrutura computacional disponível atualmente nos permitiu ver um crescimento contínuo do número de repositórios de dados de pesquisa.

Com tantas funções distintas, os repositórios tornaram-se fundamentais para o desenvolvimento de quaisquer pesquisas. Recomenda-se a publicação completa dos resultados e material suplementar em repositório de acesso aberto, garantido por uma instituição acadêmica, sociedade científica ou similar.

Nesse cenário, voltado ao constante compartilhamento de dados científicos, os repositórios de acesso aberto fizeram-se presentes no cotidiano das universidades e centros de pesquisa, proporcionando a visibilidade e acessibilidade aos resultados dos estudos dos pesquisadores dessas instituições. Eles têm sido "sistematicamente propostos pela literatura" como a ferramenta adequada para o compartilhamento dos dados científicos, principalmente pela sua interoperabilidade, arquivamento seguro e recuperação eficiente (PAGANINE; AMARO, 2020, p. 177).

Se por um lado temos os 'repositórios digitais' definidos como "bases de dados on-line que reúnem de maneira organizada a produção científica de uma instituição ou área temática" (INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2018), por outro, temos de forma mais específica os repositórios de dados científicos abertos, como sendo:

> Infraestruturas de base de dados desenvolvidas para apoiar todo o ciclo da gestão de dados de pesquisa, incluindo as ações mais dinâmicas e contundentes sobre os dados, que coletivamente são chamadas de curadoria de dados de pesquisa, que visam adicionar

> valor aos dados, avaliando, formatando, agregando e derivando novos dados (SAYÃO; SALES, 2016, p. 96).

Como é evidente, há diferenças notáveis entre os dois conceitos, especialmente com relação aos conteúdos que ambos armazenam. E quanto aos seus tipos não é dissemelhante, pois os repositórios digitais de dados científicos podem ser divididos em quatro categorias segundo Sayão e Sales (2016), sendo estas: repositórios institucionais de dados científicos, repositórios disciplinares de dados científicos, repositórios multidisciplinares de dados científicos e repositórios de dados científicos orientados por projetos.

O repositório institucional de dados científicos é caracterizado “por ser gerenciado e funcionar no âmbito de uma instituição acadêmica, como universidades ou institutos de pesquisa, e são voltados para arquivar dados que são, geralmente, provenientes unicamente das atividades acadêmicas dessas instituições” (SAYÃO; SALES, 2016, p. 101).

Como exemplo tem-se a plataforma CarpeDIEN do Instituto de Engenharia Nuclear (IEN/CNEN), um dos mais conhecidos no país, que além de armazenar dados científicos, reúne teses, dissertações, livros, entre outros.

Os repositórios temáticos de dados científicos são voltados para o arquivamento de domínios específicos de pesquisa, ou seja, uma área ou subárea específica do conhecimento como ciências ambientais ou astronomia. Além disso, eles também podem se orientar para tipos particulares de dados, como é o caso da BioModels Database, um repositório direcionado para o arquivamento, descoberta e intercâmbio de modelos computacionais na área da biologia. Outro exemplo marcante que caracteriza muito bem a categoria é o GenBank, uma base de dados publicamente disponível sobre sequências de DNA e suas traduções de proteínas (SAYÃO; SALES, 2016).

Quanto aos repositórios multidisciplinares de dados científicos, eles “reúnem coleções de dados coletados ou gerados por atividades de pesquisa em várias áreas de conhecimento. Conforme já observado, uma grande parcela dos repositórios

institucionais vinculados às universidades — pela natureza multidisciplinar dessas instituições — recai nessa categoria também" (SAYÃO; SALES, 2016, p. 103).

Podem ser classificados como pertencentes a ambas as categorias. Como exemplo, há todos os repositórios que serão analisados nesta pesquisa: Zenodo, Figshare, Dataverse, Dryad, DataHub e DANS.

Os repositórios de dados científicos orientados por projetos, como o próprio nome sugere, referem-se a coleções de dados que são resultados de projetos de pesquisa ou resoluções de problemas específicos. Há poucos exemplos para esse tipo, porém pode-se citar o The Scientific Drilling Database (SDDB), "que oferece dados de perfuração, abertos e reusáveis, que são criados no âmbito do Scientific Continental Drilling Program" (SAYÃO; SALES, 2016, p. 104).

Com relação à escolha do modo como serão feitos os compartilhamentos dos dados científicos abertos, segundo Lecardelli (2020, p. 104):

> As ferramentas e infraestruturas, repositórios, padrões de dados e metadados, linguagens, ontologias, e demais adotadas nos processos de curadoria e gestão de dados científicos, devem ser selecionadas conforme as áreas de domínio devido suas características e especificidades.

Para além de uma escolha voltada para as áreas de conhecimento dos dados que serão compartilhados, Costa (2017) diz que há duas grandes vertentes para levar em consideração, sendo a primeira relacionada às questões de infraestrutura tecnológica que permita um compartilhamento facilitado de dados e a segunda a respeito da preservação desses dados ao longo do tempo, levantando questões de armazenamento, regras de acesso e reutilização.

Há diversas opções de repositórios de dados científicos abertos, e cada uma delas possui funcionalidades e características que por vezes diferem umas das outras. Mesmo possuindo a mesma premissa de compartilhamento de dados resultantes de pesquisa científica, aspectos relacionados a direitos autorais, inserção de metadados, possibilidade de interação com outros pesquisadores dentro

da plataforma, segurança e infraestrutura, entre tantas outras, são particulares e fazem com que cada repositório possua seu diferencial.

Em 2020, como resultado do evento NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse, um grupo — que inclui pesquisadores e representantes de repositórios de dados — desenvolveu e publicou um *dataset* em que apresenta um quadro comparativo entre sete repositórios de dados científicos (STALL *et al.*, 2020).

Com o intuito de reunir as características de alguns repositórios de dados científicos abertos multidisciplinares, este estudo apresenta algumas categorias de análise e disponibiliza o mapeamento feito. Toda a fundamentação teórica priorizou apresentar, de forma equiparável, a situação do volume crescente de dados científicos e a importância do compartilhamento desses dados para então partir para perspectivas relacionadas à escolha de um repositório que atenda a essa necessidade.

3 OPÇÕES METODOLÓGICAS

A pesquisa é de natureza aplicada, caracterizada como exploratória, ao passo que buscamos nos sites de cada um dos repositórios investigados as informações descritas nas categorias previamente definidas.

Para a definição dos repositórios a serem analisados, partiu-se de artigo publicado pela Agência USP de Gestão da Informação Acadêmica ([2021]) que aborda o tema de repositórios de dados e cita seis alternativas para submeter dados de pesquisa. São eles: DANS, DataHub, Dataverse, Dryad, Figshare e Zenodo.

A partir de informações coletadas nos sites oficiais de cada um desses repositórios, entre abril e maio de 2021, foram observadas características referentes às seguintes categorias de análise: fundação e responsabilidade pelo mantenimento do repositório; idiomas da plataforma e dos *datasets* aceitos, mídias sociais e possibilidade de compartilhamento direto de conteúdo; tipos e formatos de arquivos aceitos, limites de uso, curadoria de conteúdo; identificadores persistentes de

conteúdo, de autores e informações acerca de financiamento de pesquisa; controle de versões; métricas de uso; integração com outros aplicativos e custos.

Como a fonte do estudo foram os sites institucionais dos repositórios, as análises foram feitas, sobretudo a partir de informações disponibilizadas nas abas 'Sobre' e 'Perguntas Frequentes'. Quando da inexistência dessas abas ou impossibilidade de respostas a partir delas, foram feitos levantamentos em todo o conteúdo do site.

Embora não tenham sido realizados testes de preenchimento com *datasets* experimentais, houve necessidade de realização de cadastro em repositórios, a fim de obter maiores informações sobre a visualização dos campos de preenchimento.

4 CARACTERÍSTICAS DOS REPOSITÓRIOS ANALISADOS

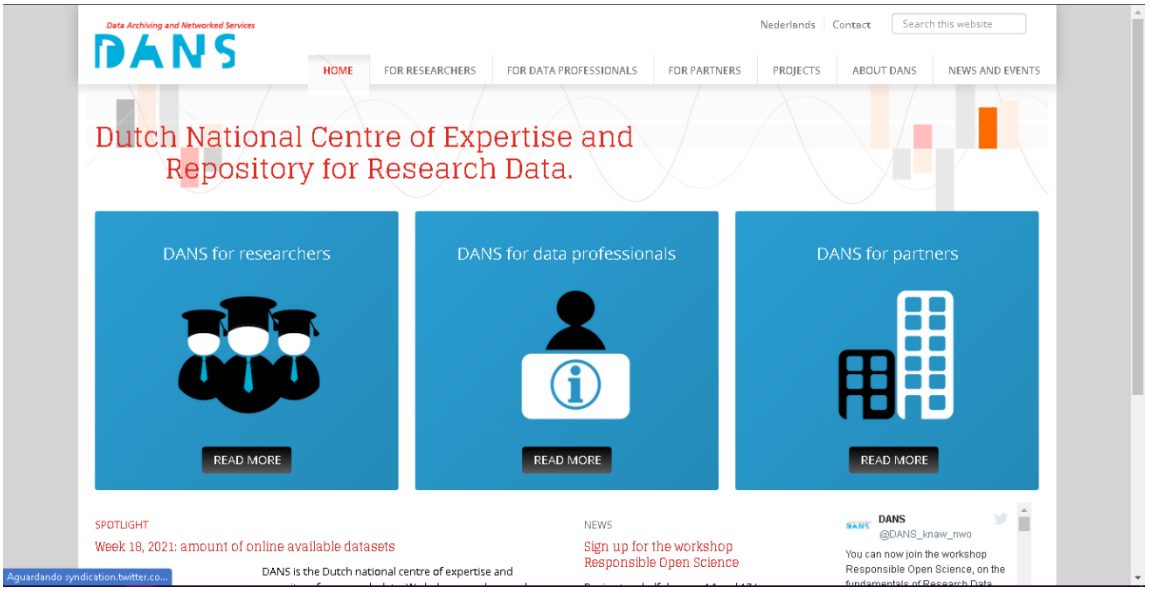
Atualmente, existem diversos repositórios multidisciplinares que abrigam dados abertos científicos e podem ser adotados como de uso comum para instituições de ensino superior. Gratuitos ou pagos, todos eles possuem uma estrutura e propósito similares, mas que, no entanto, abarcam demais características individuais que podem ou não convergir com o que cada instituição necessita, baseado em seu perfil de pesquisadores e dados científicos gerados.

Abaixo, estão descritas as características dos seis repositórios multidisciplinares selecionados pelo estudo, conforme as categorias propostas para análise.

4.1 DANS-EASY

Acrônimo de Data Archiving and Networked Services, o DANS é mantido pela KNAW (Royal Netherlands Academy of Arts and Sciences) e pelo NWO (Dutch Research Council). Fundado em 2005, o projeto conta com inúmeras possibilidades de serviços relacionados a dados, sendo um deles o EASY, repositório de dados científicos abertos. O repositório pode aparecer citado pelo nome de EASY, DANS e DANS-EASY, sendo este último o nome oficial adotado neste estudo.

Figura 1 – Tela inicial do repositório DANS-EASY



Fonte: DANS (2021).

De todos os repositórios analisados, é o único que apresenta dois idiomas diferentes de acesso ao site: inglês e holandês. Não há restrição de idiomas quanto à publicação de *datasets*, entretanto, todo conteúdo a ser publicado passa por um processo de curadoria pela equipe interna antes de sua disponibilização.

Ao publicar os *datasets*, podem ser atribuídas licenças *Creative Commons*. É atribuído automaticamente o identificador persistente DOI. Os formatos aceitos variam para cada tipo de documento, e no site há uma lista elencando quais são as extensões de arquivos preferidas por eles e quais devem ser evitadas, se possível. Após a publicação dos *datasets*, não é possível atualizar com novas versões. Caso seja necessária alguma alteração, será preciso entrar em contato com a plataforma.

É possível identificar os autores por meio de seu código ORCID, e informações acerca de financiamentos da pesquisa também estão disponíveis, no entanto esses itens não são obrigatórios na descrição. Todos os *datasets* publicados geram uma referência em formato APA. Além disso, não há possibilidade de compartilhamento desses *datasets* para mídias sociais.

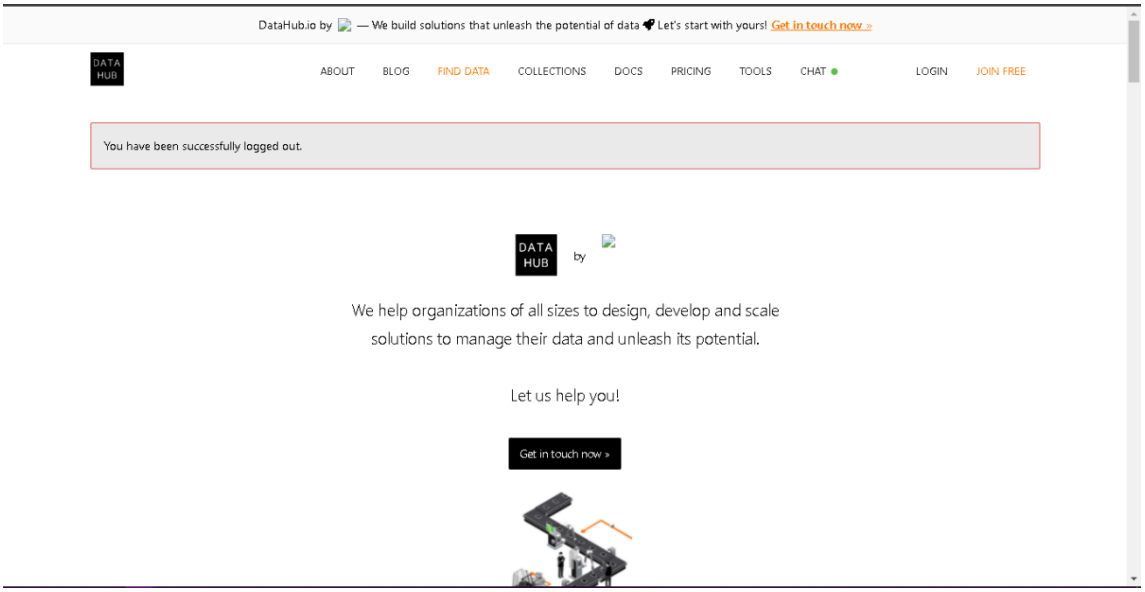
Para pesquisadores individuais com menos de 100 GB de dados publicados nos repositórios, a disponibilização de dados é gratuita. Para além disso, o

DANS-EASY não tem taxas fixas para o depósito de dados e o preço é determinado de acordo com diferentes fatores, sendo necessário entrar em contato com a plataforma.

4.2 DataHub

Fundado por Rufus Pollock e Adam Kariv, o DataHub, teve início como um projeto da Datopian e da Open Knowledge International e foi sendo aperfeiçoado até chegar ao repositório de dados abertos dos dias atuais.

Figura 2 – Tela inicial do repositório DataHub



Fonte: DataHub (2021).

Os *datasets* publicados podem ser atualizados de acordo com a periodicidade necessária, e os usuários podem interagir e conversar diretamente com os criadores por meio de uma comunidade oficial na plataforma Discord.

Há taxas para disponibilização de dados e é possível obter orçamentos para instituições organizacionais sem fins lucrativos e *startups*, no entanto existe um plano gratuito inicial, denominado *Standard*, que é oferecido pelo Datahub para publicação de dados sem custo.

O *software* aplicativo de publicação de dados, até o momento da coleta de dados para a pesquisa, estava disponível somente para instalação em dispositivos com sistema MacOS. Sempre que houver atualizações ou disponibilidade para novos sistemas, a equipe do repositório informa sobre a brevidade dessas ações em seu próprio site.

4.3 Dataverse

Embora o Dataverse configure-se como um software que possibilita a implantação de repositórios de dados e, portanto, não corresponda as mesmas características dos demais repositórios descritos neste estudo, sua permanência na análise se deu pela opção metodológica, descrita anteriormente, onde foram utilizados os indicados pela Agência USP de Gestão da Informação Acadêmica [(2021]) identificados como alternativas possíveis para a publicação de dados científicos.

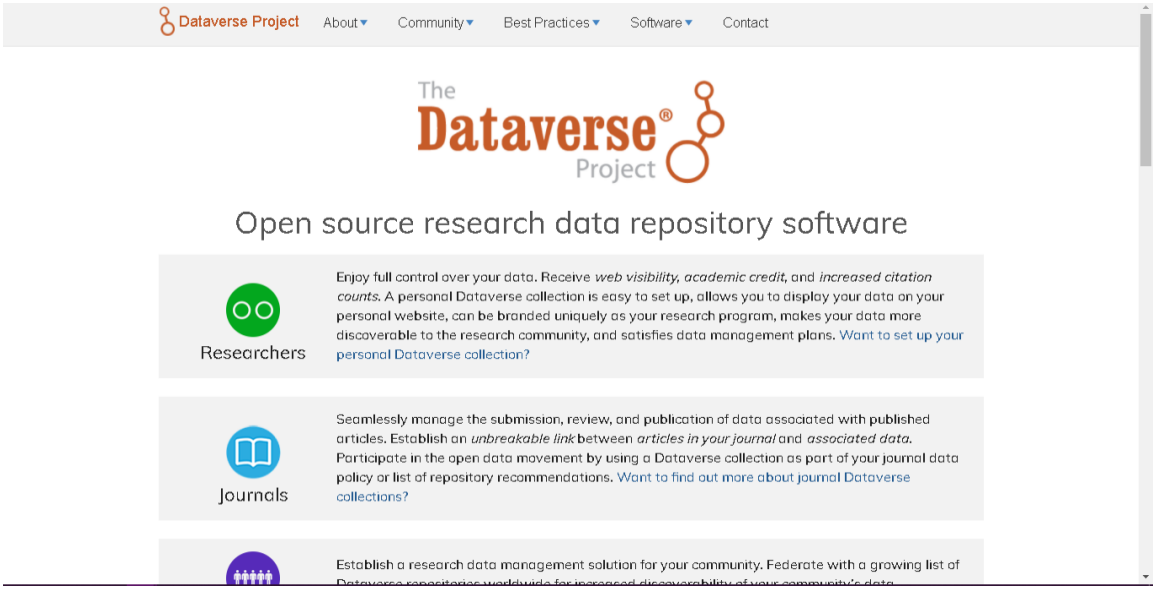
Diferentemente dos demais repositórios — onde há um serviço de depósito de dados pronto para ser utilizado pelos pesquisadores— o Dataverse possibilita a implantação de um repositório, que tem por nome ‘Coleção Dataverse’ (DATAVERSE, 2021). Esse serviço é similar ao disponibilizado pelo DSpace. Detalhes acerca desse tipo de solução tecnológica no contexto brasileiro podem ser encontrados em estudos recentes como o de Rocha *et al*. (2021).

Na ocasião da análise, utilizamos como base a combinação das informações gerais contidas na comunidade Dataverse e indicações do Harvard Dataverse Repository que é a principal coleção descrita pelo site e está aberta a todos os pesquisadores, independente das áreas do conhecimento em que atuam.

Fundado por Gary King e mantido pela Harvard's Institute for Quantitative Social Science (IQSS) e outros colaboradores, o repositório Harvard-Dataverse pode suportar publicações — livros, artigos, documentos de conferências, teses, dissertações, relatórios, entre outros — em todos os formatos com até 2.5 GB por

arquivo. Cada item publicado gera um DOI e um Handle fornecido pelo Dataverse, permitindo o controle entre as versões antigas e atualizadas.

Figura 3 – Tela inicial do Dataverse



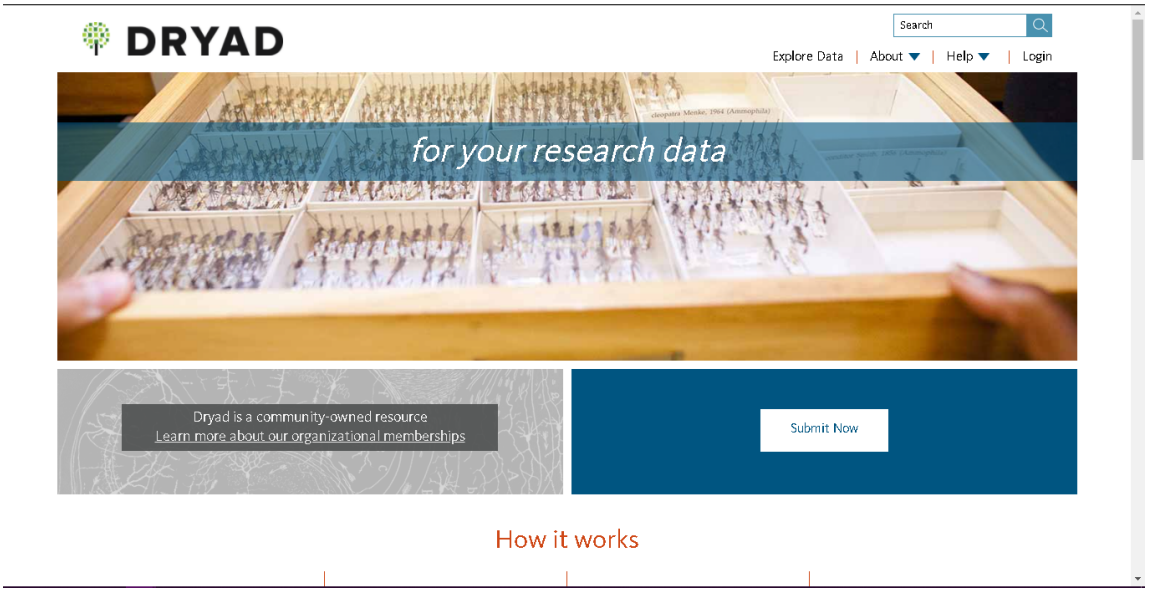
Fonte: Dataverse (2021).

Também possui integração com diversos aplicativos: Two Ravens, World Map, Open Science Framework, Open Journal Systekms, Dataverse Package for R on rOpenSci Project, EZID, DataCite, SHARE, Piwik, Rspace e Dropbox. Além disso, dentro do repositório é disponibilizado o acesso às comunidades para encontros, chamadas, entre outros tipos de interação. Porém, apesar dessas funcionalidades, não é possível ser um usuário-administrador do site sem ter a assinatura do Open Scholar ou a Harvard Key, que são serviços pagos. Ser um usuário-administrador permite um maior armazenamento de dados, opções de curadoria, capacitações de equipe gratuitas, entre outras funções.

4.4 Dryad

O Dryad é um repositório de dados abertos de pesquisa desenvolvido pelo Massachussets Institute of Technology (MIT) e Hewlett-Packard em 2009. Dez anos depois, em 2019 ele se fundiu com outro serviço de publicação de dados, o Dash.

Figura 4 – Tela inicial do repositório Dryad



Fonte: Dryad (2021).

Aceita a publicação de todos os formatos dos mais diferentes tipos de *dataset*, desde textos e planilhas até fotografias e códigos. Entretanto, o idioma de publicação, quando houver informações textuais, deve ser obrigatoriamente em inglês e, além disso, há um limite de *upload* de 300 GB por *dataset*.

Trabalha com atribuições de licenças *Creative Commons* e possui identificadores persistentes para além do DOI, como o Crossref's Funder Registry e o Research Organization Registry (RORID). Há a possibilidade de adicionar novas versões do *dataset* publicado, o DOI permanece o mesmo e as versões anteriores podem ser consultadas dentro da seção 'Arquivo de dados'.

A publicação de todo e qualquer conjunto de dados, antes da sua disponibilização, passa por uma curadoria realizada pela equipe interna do

repositório para verificação do conteúdo, que, após aprovação, segue para o processo de publicação definitivo. Os autores devem obrigatoriamente ser identificados por meio do seu ORCID.

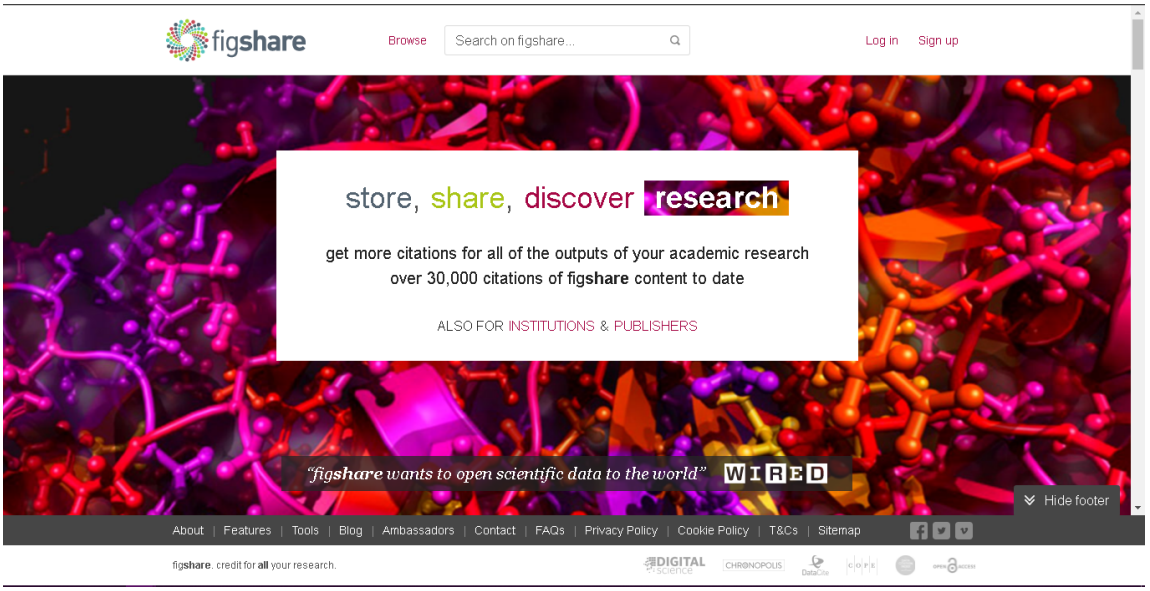
No que diz respeito a integrações e parcerias com terceiros, o Dryad possui integração com o Github e mantém uma parceria com outro repositório analisado, o Zenodo. Nessa parceria, segundo a plataforma, o foco é "potencializar os pontos fortes de cada organização: curadoria de dados na Dryad e publicação de *software* na Zenodo" (DRYAD, 2021, tradução nossa). Portanto, é possível publicar no Zenodo por meio do *upload* do Dryad, sendo essa modalidade válida somente para códigos, *scripts* e *softwares*. A publicação, feita dessa forma, estará disponível nos dois repositórios em locais distintos. Aos interessados em compartilhar dados científicos abertos no Zenodo e no Dryad simultaneamente, em ambos os sites dos repositórios há um passo a passo de como esse procedimento deve ser realizado.

De acordo com as informações do site, as taxas de publicação de dados (DPCs) custam a partir de US$ 120. A organização concede isenções para pesquisadores provenientes de países com economias tidas como de baixa renda ou média-baixa de acordo com a classificação do Banco Mundial.

4.5 Figshare

Criado em 2012, o Figshare foi desenvolvido por Mark Hahnel e é mantido pela Digital Science, uma empresa britânica que concentra seus investimentos estratégicos em empresas iniciantes que dão suporte ao ciclo de vida da pesquisa. Esse repositório aceita materiais em todos os formatos com até 5 GB de limite, englobando itens como figuras, mídias, *datasets*, pôsteres, artigos, apresentações, teses, softwares, pesquisas on-line, *pré-prints*, livros, contribuições de conferências, entre outros.

Figura 5 – Tela inicial do repositório Figshare



Fonte: Figshare (2021).

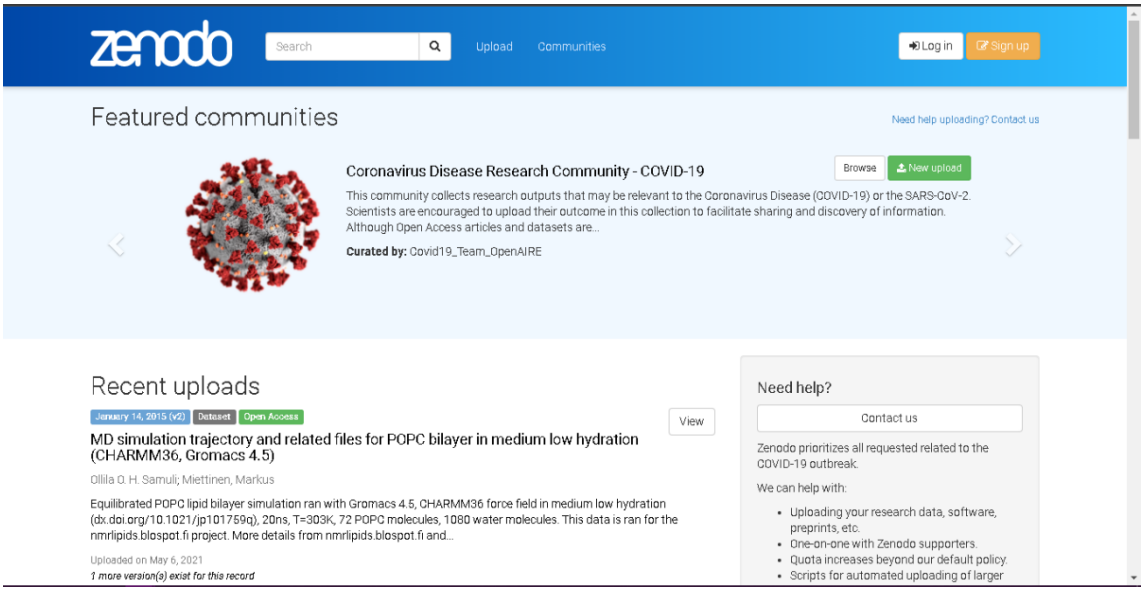
Possui integração com diversos aplicativos, entre eles: Github, Excel, Symplectic Elements (software de gerenciamento de informações acadêmicas), Binder, Elsevier Pure, Rspace, Overleaf, Open Science Framework (OSF), ImpactStory e labfolder; além da possibilidade de compartilhamento no Twitter, Facebook, Vimeo e e-mail.

Todos os itens publicados também recebem um DOI atribuído pelo Figshare, com suporte de controle de versões de todos os dados publicamente disponíveis. Quanto aos tipos de licenças aceitos, além das diversas versões do *Creative Commons*, são admitidos o MIT, GPL e Apache.

4.6 Zenodo

Repositório desenvolvido pela European Organization for Nuclear Research (CERN), o Zenodo está integrado ao projeto OpenAIRE, que tem por objetivo geral apoiar a implementação do Acesso Aberto por toda a Europa. Seu nome é derivado de Zenodotus, o primeiro bibliotecário da Antiga Biblioteca de Alexandria e pai do primeiro uso registrado de metadados.

Figura 6 – Tela inicial do repositório Zenodo



Fonte: Zenodo (2021).

São aceitos dados em quaisquer formatos, aceitando materiais como pôsteres, apresentações, conjuntos de dados, imagens, softwares, vídeos, áudios e materiais interativos como aulas, com limite de 50 GB por *dataset* (para arquivos maiores é necessário entrar em contato com os gestores ou com a equipe do repositório). Além disso, é permitido que cada usuário crie sua própria comunidade dentro do repositório, dando autonomia ao criador para gerenciá-la como quiser, gerindo ou não uma curadoria de conteúdo ao aceitar e recusar os *uploads* enviados a ela, por exemplo.

Cada item publicado recebe um DOI (Digital Object Identifier) fornecido pelo Zenodo, com a possibilidade de pequenas modificações nos metadados de arquivos recentes (publicados há menos de uma semana).

Outras funcionalidades incluem o acesso ao número de visualizações e downloads do item publicado, a compatibilidade com diversas versões da licença *Creative Commons*, integração com o Github e o compartilhamento em mídias sociais como o Twitter e Facebook.

4.7 Disponibilização e análise do mapeamento

A partir da premissa de disponibilizar em acesso aberto o conjunto de dados gerados a partir deste estudo, em cada portal eletrônico (site) dos seis repositórios pesquisados foi feita uma análise caracterizando cada um deles. A partir disso, esse mapeamento foi representado por meio de um quadro dando origem a um *dataset* em formato .pdf, que se encontra disponível no repositório digital multidisciplinar de acesso aberto Zenodo, no endereço eletrônico http://doi.org/10.5281/zenodo.5033600. A análise geral dos documentos (sites) gerou uma estrutura que apresenta o mapeamento dos repositórios analisados a partir das categorias preestabelecidas.

As características descritas em cada uma das categorias de análise foram disponibilizadas para visualização dos dados, possibilitando aos pesquisadores verificar as diferentes características de cada repositório de acordo com cada categoria proposta para análise.

Após caracterizar e mapear os repositórios analisados neste estudo, foi possível notar as possibilidades de escolha no momento de optar por um repositório para depositar os dados científicos de pesquisa, bem como a dimensão e variedade de dados que já estão disponíveis para a consulta dos pesquisadores ao redor do mundo.

Apesar das facilidades que um repositório traz para as instituições de pesquisa, seu acesso para a comunidade brasileira ainda é um pouco limitado pela barreira linguística, pois todos os repositórios descritos têm o inglês como o idioma da interface. Além disso, também houve dificuldade para obtenção de alguns dados, principalmente no que se refere ao ano de criação do repositório, idiomas de publicação aceitos e identificação de financiamento de pesquisa.

5 CONSIDERAÇÕES FINAIS

O conhecimento acerca de repositórios de dados científicos cabe não só aos pesquisadores, mas principalmente às instituições, para que elas saibam, entre as opções disponíveis, aquela que melhor se adapta à sua estrutura e para que possam desenvolver políticas institucionais que incentivem a disponibilização de dados científicos dentro das premissas estabelecidas pela Ciência Aberta (Open Science). Um padrão ou recomendação pode não só evitar a dispersão de tais dados como também criar uma cultura organizacional, facilitando posterior levantamento e análise dos dados gerados e produzidos por cada instituição ou campo temático.

No que diz respeito a escolha e utilização, ou não, dos repositórios analisados, compreendemos que o uso de repositórios não institucionais pode ser considerado do ponto de vista de suas vantagens e desvantagens. Se por um lado utilizar repositórios não institucionais impede que a instituição possua gerência total sobre ele, por outro lado, optar por um repositório deste tipo representa contenção de recursos financeiros e operacionais que seriam utilizados para manter a robusta infraestrutura que um repositório de dados científicos exige, o que pode ser um ponto crucial na escolha. Sobretudo, atualmente, onde o orçamento das universidades vem sofrendo um enxugamento orçamentário constante.

As limitações do estudo estão relacionadas principalmente às informações acerca da segurança e infraestrutura, assim como metadados e mecanismos de coleta de dados, sendo, portanto, fortes indicações para estudos futuros para que discutam o tema com um viés voltado à estrutura e equipe de TI como um dos fatores determinantes na seleção de um repositório para uso comum de uma instituição.

Nesse sentido, é importante lembrar que antes da escolha de repositórios multidisciplinares, os pesquisadores devem determinar se existe um repositório temático apropriado para seus dados de pesquisa e analisar se há recomendações de seu campo de pesquisa sobre alguma direção.

Como lembrado por Stall *et al.* (2020), os pesquisadores precisam cumprir os requisitos de seu campo de pesquisa, sua agência financiadora, país de origem, editora, entre outros, para garantir que o melhor repositório seja selecionado, de forma individualizada.

Por fim, esperarmos ter contribuído com a caracterização e mapeamento dos repositórios de dados científicos abertos como alternativa para a tomada de decisão institucional no campo da gestão de dados científicos.

Abstract: The sharing scientific data makes it possible to accelerate the construction of new knowledge, increasing the efficiency of resources, expanding the possibilities for reuse and reproducibility of research. In order to present a set of information with characteristics of open scientific data repositories, it was defined as an objective for this study to characterize and map open scientific data repositories as an alternative for institutional decision-making in the field of scientific data management. An exploratory-descriptive, qualitative study was carried out and, in relation to technical procedures, a bibliographic-documentary survey was carried out, where six open scientific data repositories were analyzed: DANS, DataHub, Harvard-Dataverse, Dryad, Figshare and Zenodo. Based on information collected from the official websites of each of these repositories between April and May 2021, and the following characteristics were observed: foundation and responsibility for maintaining the repository; platform languages and accepted datasets, social media and the possibility of direct content sharing; file types and formats accepted, usage limits, content curation; persistent identifiers of content, authors, and research funding information; version control; usage metrics; integration with other applications and costs. The analysis of the sites generated a table, available in the Zenodo repository, with the description of the data of each repository from the categories. The results presented are intended to assist researchers in their choice of an open scientific data repository that allows the sharing of research data.

Keywords: Scientific data repositories. Open scientific data. Multidisciplinary repositories.

REFERÊNCIAS

AGÊNCIA USP DE GESTÃO DA INFORMAÇÃO ACADÊMICA. **Repositórios de Dados**. São Paulo, [2021]. Disponível em: https://www.aguia.usp.br/apoio-pesquisador/dados-pesquisa/lista-repositorios-dados-pesquisa/. Acesso em: 07 jun. 2021.

CIUFFO, Leandro *et al.* Acesso aberto a dados de pesquisa. *In*: WORKSHOP RNP, 18., Belém, 2017. Disponível em: http://wrnp.rnp.br/sites/wrnp2017/files/11_wrnp2017_cartaz_aadp_design.pdf. Acesso em: 12 mar. 2021.

COSTA, Maíra Murrieta. **Diretrizes para uma política de gestão de dados científicos no Brasil**. 2017. Tese (Doutorado em Ciência da Informação) — Universidade de Brasília, Brasília, 2017. Disponível em: https://repositorio.unb.br/handle/10482/24895. Acesso em 12 out. 2021.

COVID-19 DATA SHARING/BR. [S. l., 2021]. **Repositório de dados científicos.** Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/. Acesso em: 12 out. 2021.

DANS-EASY. **Data Archiving and Networked Services.** [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://dans.knaw.nl/en. Acesso em: 06 maio 2021.

DATAHUB. **Datahub**. [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://datahub.io. Acesso em 06 maio 2021.

DATAVERSE. **The Dataverse project.** [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://dataverse.org. Acesso em: 06 maio 2021.

DRYAD. **Dryad Digital Repository.** [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://datadryad.org/stash. Acesso em: 06 maio 2021.

EUROPEAN COMMISSION. **Open Innovation, Open Science, Open to the World**: a vision for Europe. Luxembourg: Publications Office of the European Union, 2016**.** Disponível em: https://op.europa.eu/en/publication-detail/-/publication/3213b335-1cbc-11e6-ba9a-01aa75ed71a1. Acesso em: 16 mar. 2021.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. **Repositório COVID-19 Data Sharing/BR viabiliza descobertas nas áreas da saúde e da computação**. São Paulo, 2021. Disponível em: https://agencia.fapesp.br/repositorio-covid-19-data-sharing-br-viabiliza-descobertas-nas-areas-da-saude-e-da-computacao/36205/. Acesso em: 12 out. 2021.

FERREIRA, Valdinéia Barreto. E-science. *In*: **E-science e políticas públicas para ciência, tecnologia e inovação no Brasil**. Salvador: EDUFBA, 2018. p. 13-30. Disponível em: https://doi.org/10.7476/9788523218652.0003. Acesso em: 29 out. 2021.

FIENBERG, Stephen Elliott; MARTIN, Margaret Elizabeth; STRAF, Miron Lowel (org.). **Sharing Research Data**. Washington, DC.: National Academy Press, 1985. 240 p. Disponível em: https://doi.org/10.17226/2033. Acesso em: 20 out. 2021.

FIGSHARE. Figshare. [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://figshare.com. Acesso em: 06 maio 2021.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Repositórios Digitais**. 2018. Disponível em: http://sitehistorico.ibict.br/informacao-para-ciencia-tecnologia-e-inovacao%20/repositorios-digitais. Acesso em: 22 mar. 2021.

LECARDELLI, Jane. **Dados científicos abertos em agências de fomento à pesquisa:** cenário dos Planos de Gestão de Dados (PGD) e Princípios **FAIR**. 2020. Dissertação - Universidade do Estado de Santa Catarina, Centro de Ciências Humanas e da Educação, Programa de Pós-Graduação em Gestão de Unidades de Informação, Florianópolis, 2020. Disponível em: https://sistemabu.udesc.br/pergamumweb/vinculos/00008b/00008bdf.pdf. Acesso em: 10 dez. 2021.

MONTEIRO, Gabriela. **Mapeamento e análise das políticas institucionais de financiadores da pesquisa brasileira**: cenário dos dados científicos abertos. 2019. 146 f. Dissertação - Universidade do Estado de Santa Catarina, Centro de Ciências Humanas e da Educação, Programa de Pós-Graduação em Gestão de Unidades de Informação, Florianópolis, 2019. Disponível em: https://sistemabu.udesc.br/pergamumweb/vinculos/00007a/00007ace.pdf. Acesso em: 10 dez. 2021.

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. **Principles and guidelines for access to research data from public funding**. Paris: OECID, 2007. Disponível em: http://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 12 jun. 2021.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **¿Puede la "Ciencia Abierta" acelerar la búsqueda de una vacuna contra el COVID-19? Cinco cosas que debes saber**. 2020. Disponível em: https://news.un.org/es/story/2020/11/1483842. Acesso em: 16 mar. 2021.

ORTIZ-REPISO, Virginia; HERNÁNDEZ-PEREZ, Antonio. **Curatore+**: custodia y gestión digital para la reutilización de datos abiertos de investigación. [Projeto de Pesquisa]. Universidad Carlos III de Madrid. Madrid (ES), 2017.

PAGANINE, Lucas Nóbrega; AMARO, Bianca. Características dos repositórios de dados científicos no Brasil. **Biblos**, v. 34, n. 1, p. 176-188, 2020. Disponível em: https://doi.org/10.14295/biblos.v34i1.11132. Acesso em: 29 mar. 2021.

ROCHA, Rafael Port da *et al.* Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 17, p. 1-25, 2021. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1572/1261. Acesso em: 14 out. 2021.

SAYÃO, Luis Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa**. Informação & Informação**, v. 21, n. 2, p. 90-115, 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 29 mar. 2021.

SILVA, Maria Helena Ferreira Xavier da *et al*. Competências dos bibliotecários na gestão dos dados de pesquisa. **Ciência da Informação**, v. 48, n. 3, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/136517. Acesso em: 12 mar. 2021.

STALL, Shelley *et al.* **Generalist Repository Comparison Chart**. 2020. Disponível em: https://doi.org/10.5281/ZENODO.3946720. Acesso em: 17 jun. 2021.

ZENODO. Zenodo. [*S. l.*], 2021. Repositório de dados científicos. Disponível em: https://zenodo.org. Acesso em: 06 maio 2021.