30º CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO
25 a 29 de novembro 2024
Bibliotecas Fortes: Sociedade Democrática
Recife, PE

Eixo 6 – O mundo digital: apropriação e desafios

Modalidade: trabalho completo

Características dos repositórios de dados de pesquisa brasileiros: uma análise a partir do diretório re3data

Characteristics of Brazilian research data repositories: an analysis based on re3data

Letícia Guarany Bonetti – Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Tatyane Guedes Martins da Silva – Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Rene Faustino Gabriel Junior – Universidade Federal do Rio Grande do Sul (UFRGS)

Blena Estevam dos Santos – Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Millena Cordeiro Matos de Lima – Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Priscila Machado Borges Sena – Universidade Federal do Rio Grande do Sul (UFRGS)

Resumo: Os dados de pesquisa são um insumo valioso para a ciência, e por isso fala-se na importância da sua gestão em repositórios de dados. Esta pesquisa tem como objetivo identificar as principais características dos repositórios de dados de pesquisa no Brasil, comparando os resultados com o estudo realizado por Paganini e Amaro em 2020. Trata-se de uma pesquisa descritiva que utiliza o diretório re3data como fonte de dados. Os resultados indicam um crescimento linear no número de repositórios brasileiros cadastrados no diretório, apresentando características semelhantes às da pesquisa anterior. No entanto, novas características foram observadas, como a predominância do Dataverse.

Palavras-chave: Repositório de dados. Re3data. Dado de pesquisa. Dados Abertos.

Abstract: Research data is a valuable input for science, which is why there is talk of the importance of managing it in data repositories. This research aims to identify the main characteristics of research data repositories in Brazil, comparing the results with the study carried out by Paganini and Amaro in 2020. This is a descriptive research that uses the re3data directory as a data source. The results indicate a linear growth in the number of Brazilian repositories registered in the directory, presenting characteristics similar to

those of the previous research. However, new characteristics were observed, such as the predominance of Dataverse.

Keywords: Data repository. Re3data. Research data. Open Data.

1 INTRODUÇÃO

No contexto das pesquisas científicas, um insumo fundamental são os dados coletados, e uma forma de potencializar seus benefícios é preservá-los em um repositório. Instituições e grupos de pesquisa têm criado repositórios institucionais e temáticos motivados principalmente pelo desenvolvimento de recursos como as Tecnologias da Informação e Comunicação (TIC). Essa evolução e sofisticação das tecnologias leva a cenários como o *Big Data*[1], caracterizado pelo grande volume de dados e pelo uso de computação intensiva.

Entende-se que os dados são insumos valiosos para a comunidade científica e sua definição pode variar de acordo com seu domínio. Os dados de pesquisa são heterogêneos e podem ser "[...] números, figuras, vídeos, softwares; com diferentes níveis de agregação e de processamento, como dados crus ou primários, dados intermediários e dados processados e integrados; e em diferentes formatos de arquivos" (Sayão; Sales, 2016, p. 94).

É preciso ter em mente que muitos dos dados coletados por pesquisadores não podem ser substituídos em casos de perda ou destruição como, por exemplo, dados de eventos climáticos. Desta forma, os dados são singulares, importantes para registro e validação de estudos. É nesse contexto de valorização que surgem ambientes como os repositórios de dados de pesquisa, que são centrados em gerenciar e facilitar seu fluxo, sendo a ferramenta de armazenamento e gerenciamento dos dados utilizados e produzidos nas pesquisas (Rodrigues; Dias; Lourenço, 2022).

Em suma, os repositórios de dados são infraestruturas tecnológicas projetadas para auxiliar todo o ciclo de gestão de dados de pesquisa, abrangendo as ações mais dinâmicas e impactantes sobre os dados, conhecidas coletivamente como curadoria de dados de pesquisa (Sayão; Sales, 2016). Assim, fornecem visibilidade, arquivamento,

[1] *Big Data* é um termo utilizado para um conjunto massivo de dados, que podem apresentar dificuldades de armazenamento, análise e visualização (Sagiroglu; Sinanc, 2013, tradução nossa).

recuperação, compartilhamento, reconhecimento de autoria, preservação digital, memória científica, disseminação e acesso aos dados de pesquisa, pautando-se pelos valores e princípios do movimento da Ciência Aberta (Sanchez; Vidotti; Vechiato, 2017).

Quando compartilhados em ambientes como os repositórios, os dados podem ser 1) reutilizados para novas pesquisas, economizando recursos e acelerando a inovação, e 2) reproduzidos, garantindo a integridade e a transparência do processo científico, passando a considerar também seu impacto social. O acesso igualitário, contínuo e aberto à informação e aos dados é uma premissa de suma importância para preservação do patrimônio científico e digital da humanidade, beneficiando toda a sociedade (Sayão; Sales, 2016).

Nota-se que, no contexto brasileiro, as iniciativas de repositórios de dados de pesquisa alinham-se aos princípios do *Open Data* (Dados Abertos), um dos pilares da Ciência Aberta, que preconiza que os dados sejam acessados, usados e compartilhados para qualquer finalidade, sem restrições. A exemplo, tem-se a solicitação de um "Plano de Gerenciamento de Dados" pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)[2], uma das principais agências de fomento no país, que já incentiva pesquisadores a depositarem seus dados de pesquisa em ambientes digitais confiáveis.

Apesar da sua importância nos ecossistemas da pesquisa científica brasileira e da Ciência Aberta, pesquisadoras e pesquisadores ainda encontram dificuldade na gestão de seus dados e na própria escolha de ambientes seguros para depositá-los, considerando especificidades como a área de pesquisa e a tipologia dos dados produzidos. Essa demanda levou à criação de catálogos de repositórios de dados voltados para identificação, descrição e acesso à repositórios, e um dos mais conhecidos é o *Registry of Research Data Repositories*, o re3data[3].

O diretório re3data, lançado em 2012, é um registro global de repositórios de dados de pesquisa. Ele promove uma cultura de compartilhamento, acesso e visibilidade aos dados, sendo um serviço parceiro da *DataCite*, uma organização global sem fins lucrativos que está ativamente envolvida em várias iniciativas para melhorar a disponibilidade e citação de resultados de pesquisa (Rücknagel *et al.*, 2021).

[2] Disponível em: https://fapesp.br/.
[3] Disponível em: https://www.re3data.org/

O re3data tem a função de auxiliar pesquisadores, organizações financiadoras e editores no que tange os repositórios de dados de pesquisa. Devido à sua importância global e à facilidade de consulta em seu sistema, o diretório foi escolhido como fonte de dados para este trabalho, que busca identificar as principais características dos repositórios de dados de pesquisa do Brasil.

2 PERCURSO METODOLÓGICO

Por meio de pesquisa descritiva, buscou-se identificar as principais características dos repositórios de dados de pesquisa do Brasil indexados no diretório re3data. Um estudo similar foi realizado por Paganine e Amaro (2020), que analisaram características dos repositórios como: software, licença, padrão de metadados, quantidade de itens depositados, *etc*. No estudo supracitado, o re3data continha um total de nove (9) repositórios brasileiros. Nesta pesquisa foram identificados vinte e um (21) repositórios, um acréscimo de 12 repositórios em quatro anos. Com isso em mente, este estudo tem como foco descrever as características e situação atual de desenvolvimento dos repositórios de dados de pesquisa no Brasil. A proposta é levantar as principais características dos repositórios e fazer uma comparação com o cenário de 2020, analisando as mudanças e tendências atuais. Salienta-se que algumas características apontadas neste estudo não foram analisadas no artigo citado (Paganine; Amaro, 2020), impossibilitando a comparação.

O próprio re3data oferece as informações sobre os repositórios, sendo então a fonte de coleta de dados para esta pesquisa. Já a existência de uma política de gestão foi verificada diretamente no site dos repositórios. Vale salientar que esses repositórios são cadastrados no re3data pelos seus gestores ou outros responsáveis, que preenchem o formulário de inscrição declarando as informações solicitadas sobre o repositório. Após a inscrição, a equipe do re3data analisa se os requisitos mínimos da política de dados foram cumpridos, e o repositório é indexado no diretório.

O diretório re3data permite buscas por assunto, tipo de conteúdo e país. Para este trabalho, realizou-se a busca por país no dia 18 de junho de 2024, que retornou, como já citado, um total de 21 repositórios de dados de pesquisa no Brasil. Os repositórios indexados no re3data são: 1) *WorldClim - Global Climate Data*; 2) GLOBE;

3) Deposita Dados; 4) Base de Dados Científicos da Universidade Federal do Paraná; 5) FAPESP COVID-19 *Data Sharing*/BR; 6) PPBio *Data Repository*; 7) Repositório de Dados de Pesquisa da Embrapa (Redape); 8) Repositório de Dados de Pesquisa da Unifesp; 9) IBICT *Cariniana Dataverse Network*; 10) *International Ocean Discovery Program*; 11) CEDAP *Research Data Repository*; 12) Repositório Institucional da UNESP; 13) Arca Dados; 14) *FishSounds*; 15) SciELO *Data*; 16) Repositório de Dados de Pesquisa da Unicamp (REDU); 17) Maenduar; 18) *Exploration and Production Data Bank*; 19) *Open Research Data* @PUC-Rio; 20) Repositório de Dados de Pesquisas do Instituto Federal Goiano; 21) Aleia.

Entretanto, após realizar uma consulta no site dos repositórios, constatou-se que dois deles não eram, de fato, repositórios de dados de pesquisa. O Maenduar se trata de uma comunidade dentro do Zenodo, que é um repositório multidisciplinar de Acesso Aberto. O Maenduar foi criado pelo Laboratório em Rede de Humanidades Digitais do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), não sendo, portanto, um repositório em si, mas uma comunidade dentro de um repositório internacional operado pela *European Organization for Nuclear Research* (CERN). Já o *Exploration and Production Data Bank* não se encontra mais em operação. Apesar de ainda estar indexado no re3data e ser possível consultar os metadados no diretório, não é mais possível acessar o repositório em si. Logo, ambos foram excluídos da amostra.

A amostra levantada neste trabalho não equivale ao total de repositórios de dados existentes no Brasil, mas, para fins de comparação com estudo anterior, optou-se por analisar apenas os repositórios indexados no diretório re3data, que fornece diversos dados sobre os repositórios. Desse modo, a pesquisa percorreu seis etapas:

1. Busca por país no diretório re3data (filtro: Brasil);
2. Coleta dos dados dos 21 repositórios brasileiros indexados no re3data;
3. Exportação dos dados para cada um dos 21 repositórios em uma Planilha Google;
4. Análise do site dos 21 repositórios para checar se todos correspondem, de fato, a repositórios de dados e se possuem políticas de gestão;
5. Exclusão dos repositórios que estão mapeados no re3data, mas não se configuram como repositórios de dados, totalizando uma amostra de 19 repositórios;

6. Limpeza dos dados na Planilha para a análise e geração de gráficos, além da preservação para futuro compartilhamento em um repositório de dados de pesquisa.

A partir do percurso descrito acima, obteve-se os dados que subsidiam a discussão da próxima seção. Ressalta-se que a Planilha Google citada possibilita a identificação da autoria, e por isso será disponibilizada após a avaliação por pares.

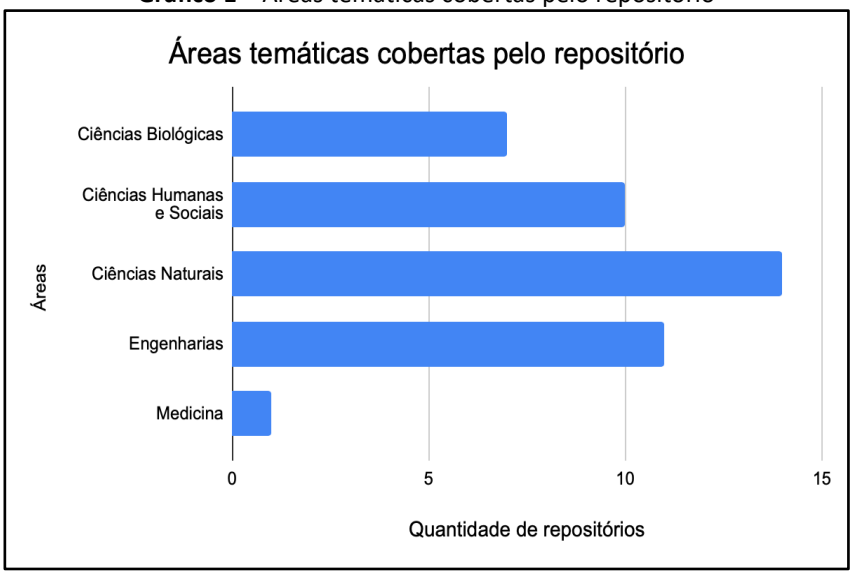
3 RESULTADOS E DISCUSSÕES

Para cada um dos 19 repositórios da amostra foram coletados os seguintes dados, que serão analisados neste trabalho: áreas temáticas; tipo de repositório (disciplinar ou institucional); tipo de acesso aos dados; software; licença; padrão de metadados; identificador persistente; versionamento; política de gestão do repositório.

3.1 Áreas temáticas

O Gráfico 1 contempla as áreas do conhecimento dos repositórios brasileiros. Optou-se por indicar apenas as áreas gerais declaradas pelos repositórios no re3data, e não as subáreas.

Gráfico 1 – Áreas temáticas cobertas pelo repositório



Fonte: Elaborada pelas autoras (2024).
Descrição: Apresenta um gráfico horizontal de cinco barras em uma só cor, contendo como título: "Áreas temáticas cobertas pelo repositório". De cima para baixo estão listadas: Ciências Biológicas, Ciências Humanas e Sociais, Ciências Naturais, Engenharias e, por último, Medicina. Cada área temática apresenta a quantidade de repositórios que ela aparece e o gráfico conta com um intervalo de contagem de cinco em cinco. No eixo vertical tem-se a legenda "áreas" e no horizontal "quantidade de repositórios"

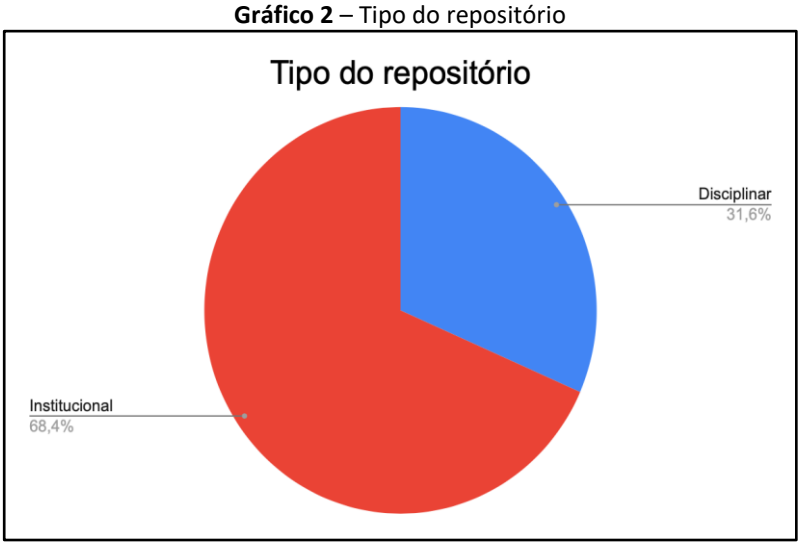
Nota-se uma predominância de cobertura dos repositórios nas áreas das Ciências Naturais, declarada por 14 repositórios, e Engenharias, declarada por 11 repositórios. Isso significa que a maioria dos repositórios brasileiros abarcam as áreas da ciência que estudam a natureza e as engenharias como um todo, como é o caso do repositório GLOBE e do Deposita Dados. Mas também é possível observar um número expressivo de repositórios nas áreas de Humanas e Sociais.

Alguns repositórios apresentam mais de uma área temática como, por exemplo, o IBICT *Cariniana Dataverse Network*, que possui conjuntos de dados das Ciências Humanas e Sociais, Ciências Naturais, Ciências Biológicas e Engenharias. Em estudo anterior realizado por Paganini e Amaro (2020, p. 178) constatou-se que "[...] os temas estão distribuídos de forma homogênea, exceto entre as Engenharias, em que aparece em apenas um repositório". Nesses quatro anos percebe-se um aumento de repositórios na área de Engenharias, que corresponde a 11 dos 19 repositórios analisados.

3.2 Tipo de repositório

Conforme Sayão e Sales (2016), os repositórios de dados podem ser divididos, de acordo com a literatura, em quatro tipos: institucionais, disciplinares, multidisciplinares e orientados por projetos. Entretanto, o re3data divide os repositórios em apenas dois tipos: institucional (*institutional*) ou disciplinar (*disciplinary*), permitindo ainda a opção "outro".

Gráfico 2 – Tipo do repositório



Fonte: Elaborada pelas autoras (2024).

Descrição: Apresenta um gráfico em formato de pizza, que está dividido em duas partes, com cores distintas. O título está descrito como “Tipo de repositório”. A fatia azul, que é a menor, representa o

repositório disciplinar, com a porcentagem de 31,6%. A segunda fatia, que é a maior, representa o repositório institucional, com a porcentagem de 68,4%.

Os repositórios institucionais costumam cumprir o papel de memória institucional e são aqueles gerenciados "[...] 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). Já os repositórios disciplinares são aqueles "[...] voltados para o arquivamento de domínios específicos de pesquisa como física de partículas ou ciências ambientais" (Sayão; Sales, 2016, p. 102).

No Gráfico 2 é possível observar que o tipo de repositório predominante no Brasil é o institucional, que corresponde a 68,4% da amostra, ou seja, 13 repositórios. Isso significa que a maioria dos repositórios são gerenciados no âmbito de uma instituição, armazenando os dados coletados pelo seu quadro funcional, que envolve pesquisadores, estudantes, professores, e a comunidade científica no geral. Alguns exemplos são o Redape, da Empresa Brasileira de Pesquisa Agropecuária (Embrapa); o Arca Dados, da Fundação Oswaldo Cruz (Fiocruz); e o Aleia, do Ibict. Os repositórios disciplinares, voltados para domínios específicos, equivalem a 31,6% da amostra, 6 repositórios, e alguns exemplos são a FAPESP COVID-19 *Data Sharing*/BR (Medicina) e o PPBio *Data Repository* (Ciências Biológicas).

3.3 Tipo de acesso aos dados

O tipo de acesso aos dados pode ser aberto, embargado ou restrito. De acordo com as informações fornecidas no re3data, todos os 19 repositórios da amostra são de acesso aberto, mesmo resultado encontrado em estudo anterior (Paganine; Amaro, 2020). Ou seja, a tendência do acesso aberto mantém-se entre os repositórios do Brasil.

Ramachandran, Bugbee e Murphy (2021) explicam que, já que os dados são essenciais para os processos científicos, há um esforço dentro do movimento da Ciência Aberta para torná-los mais abertos, incentivando seu compartilhamento e disponibilização em ambientes como os repositórios, no que é conhecido como o movimento dos Dados Abertos. De acordo com Caballero-Rivero, Sánchez-Tarragó e Santos (2019, não paginado), dados abertos "[...] são aqueles que podem ser usados livremente, reutilizados e redistribuídos por qualquer pessoa, sujeitos unicamente aos

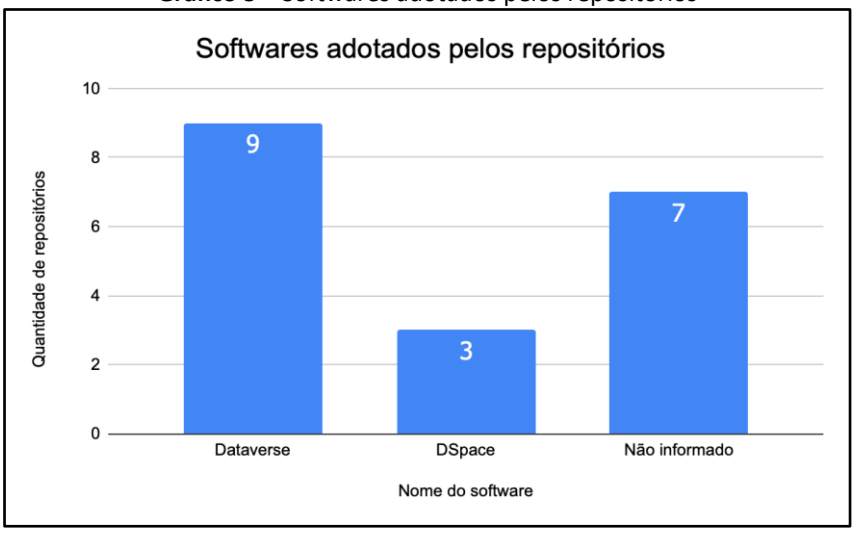
requisitos de atribuição de autoria e de compartilhamento na mesma forma em que foram obtidos".

Entretanto, é importante destacar que nem todos os dados podem ser abertos. Ramachandran, Bugbee e Murphy (2021) explicam que dados podem sofrer restrições de acesso devido a questões como a segurança nacional, a presença de dados sensíveis ou a proteção para patentes. É o que se vê em alguns casos da amostra como a FAPESP COVID-19 *Data Sharing*/BR, o PPBio *Data Repository*, o Redape e o Arca Dados, que possuem conjuntos de dados restritos ou embargados, mesmo sendo repositórios de acesso aberto. Nesse sentido adota-se a premissa "[...] tão aberto quanto possível, tão fechado quanto necessário" (*European Commission*, 2016, p. 4, tradução nossa).

3.4 Software

Dos 19 repositórios, 7 não informaram no re3data o software adotado para a construção do ambiente digital, 9 adotam o Dataverse, sendo o software predominante, e 3 adotam o DSpace, como exposto no Gráfico 3.

Gráfico 3 – Softwares adotados pelos repositórios



Fonte: Elaborada pelas autoras (2024).
Descrição: Apresenta um gráfico vertical em uma só cor, sob o título "Softwares adotados pelos repositórios". Em seu eixo vertical está sob o título "Quantidade de repositórios", e em seu eixo horizontal, "nome do software". O gráfico está dividido em um intervalo de contagem de cinco em cinco. O primeiro software apresentado é o "Dataverse", presente em 9 repositórios; em seguida, o "DSpace", em 3; e, por último, "não informado", em 7.

O Dataverse foi desenvolvido em 2006 pelo *Institute for Quantitative Social Science* da Universidade de Harvard em colaboração com profissionais de todo o mundo, por meio do *Dataverse Project* (DATAVERSE, 2020). A essência do *Dataverse Project* é

reduzir a carga de trabalho de pesquisadores e editores de dados, através da automatização, com foco nos dados de pesquisa. Sendo um software livre de código aberto, é possível modificá-lo e distribuir versões modificadas, além de personalizar o sistema para adaptá-lo às necessidades locais.

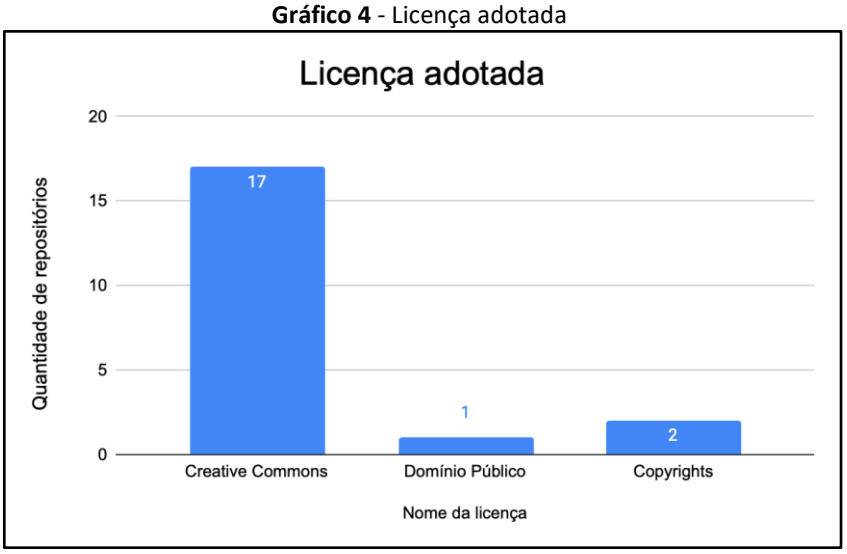
Já o DSpace nasceu de um esforço conjunto entre os desenvolvedores do *Massachusetts Institute of Technology (MIT) e Hewlett-Packard (HP)*. No presente momento, a organização *DuraSpace* desenvolve, apoia e promove a utilização do software em âmbito mundial. Assim como o Dataverse, o DSpace é um software livre de código aberto, que permite o armazenamento, gestão, preservação e a visibilidade da produção intelectual em repositórios e bibliotecas digitais. O DSpace, ao contrário do Dataverse, não foi desenvolvido com foco em dados de pesquisa, exigindo uma série de adaptações nas configurações para se adequar à representação desses objetos digitais (Rocha *et al*., 2021).

Em estudo similar realizado quatro anos atrás (Paganine; Amaro, 2020), não houve a predominância de um software específico, tendo apenas um repositório DSpace e um repositório Dataverse entre os 9 repositórios analisados pelos autores. Houve, portanto, um maior investimento no Brasil ao longo dos anos para a adoção desses dois softwares mundialmente conhecidos.

3.5 Licença

No gráfico 4 é possível visualizar as licenças adotadas pelos repositórios da amostra. A licença *Creative Commons* (CC) é a que tem mais ocorrência (17), seguido por *Copyrights* (2) e Domínio Público (1). O resultado se assemelha ao que foi levantado por Paganine e Amaro (2020) em seu estudo, em que 5 dos 9 repositórios da amostra adotavam o CC, enquanto 2 adotavam a licença *Copyrights*.

Gráfico 4 - Licença adotada



Fonte: Elaborada pelas autoras (2024).
Descrição: Apresenta um gráfico vertical em uma só cor, sob o título "Licença adotada". Em seu eixo vertical está sob o título "Quantidade de repositórios" e em seu eixo horizontal, "Nome da licença". O gráfico está dividido em um intervalo de contagem de cinco em cinco. A primeira licença é a "Creative Commons", representada com o número 17; a segunda é "Domínio Público", com 1; e, por fim, "Copyrights", com 2.

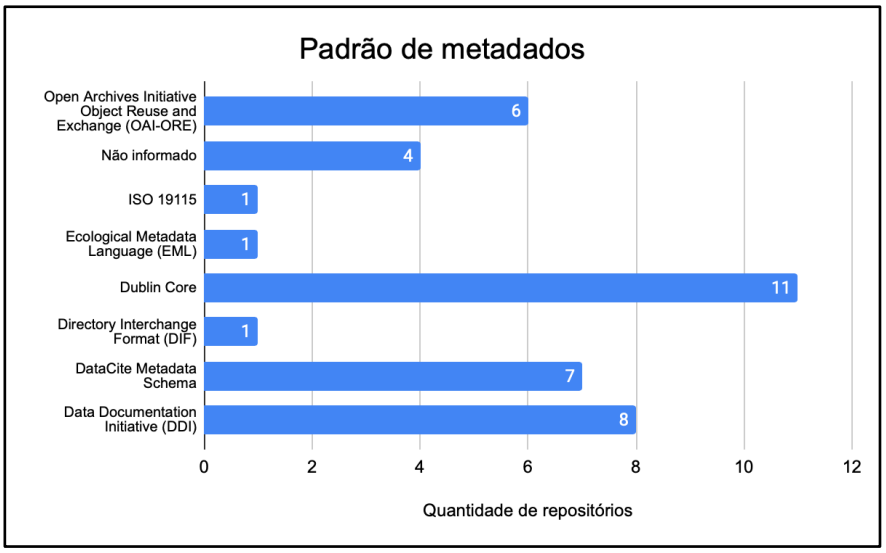
Como o *Creative Commons* é conhecido mundialmente, principalmente atrelado às iniciativas de Acesso Aberto, era esperado que a maioria dos repositórios adotassem as licenças propostas pela organização. Já o *Copyright* "[...] garante aos proprietários dos direitos econômicos sobre a obra a proteção sobre os direitos de reprodução de cópias, ou seja, o editor detém o controle sobre a reprografia, a lei garante sua proteção e o direito de cobrar pelas obras copiadas" (Reis; Rozados, 2013, p. 67). Percebe-se, portanto, que há uma predominância do uso de licenças mais permissivas nos repositórios, incentivando que os dados sejam distribuídos, editados, remixados e utilizados para criar outros trabalhos, dando os devidos créditos aos autores. Isso vai em direção do que é defendido pelo movimento da Ciência Aberta, que busca democratizar o acesso e incentivar o trabalho colaborativo entre pesquisadores.

3.6 Padrão de metadados

Conforme mostra o Gráfico 5, o padrão de metadados mais utilizado entre os repositórios da amostra é o *Dublin Core* (DC), em 11 repositórios. De acordo com Souza, Vendrusculo e Melo (2000, p. 93), o DC "[...] pode ser definido como sendo o conjunto de elementos de metadados planejado para facilitar a descrição de recursos

eletrônicos". Ele é composto de 15 elementos essenciais e tem como objetivo a simplicidade, a flexibilidade, a semântica e a interoperabilidade entre sistemas.

Gráfico 5 - Padrão de metadados



Fonte: Elaborada pelas autoras (2024).
Descrição: Apresenta um gráfico horizontal em uma só cor, sob o título "Padrão de metadados". O gráfico está dividido em um intervalo de contagem de dois em dois. Os padrões de metadados estão listados de cima para baixo na seguinte ordem: o primeiro é o "Open Archives Initiative Object Reuse and Exchange (OAI-ORE)", com 6 repositórios; em seguida, "não informado", com 4; "ISO 19115", com 1; "Ecological Metadata Language (EML), com 1; "Dublin Core", com 11; "Directory Interchange Format (DIF), com 1; "DataCite Metadata Schema", com 7; e "Data Documentation Initiative (DDI), com 8".

Apesar de não ter sido desenvolvido com foco na descrição de dados de pesquisa, o DC é um padrão amplamente adotado em repositórios no Brasil e pode ser empregado para a representação dos dados de pesquisa, sozinho ou junto com outros padrões de metadados. Rocha *et al.* (2021, p. 13) explicam que "O Dataverse, ao permitir o mapeamento de seus metadados para Dublin Core e para DataCite, dá suporte à descoberta de informações".

Em seguida, os padrões mais adotados são o *Data Documentation Initiative*[4] (DDI), em 8 repositórios, o *DataCite Metadata Schema*[5], em 7 repositórios, e o *Open Archives Initiative Object Reuse and Exchance*[6] (OAI-ORE), em 6. Ao contrário do DC, que é um padrão genérico, o DDI é voltado para dados. No estudo de Paganine e Amaro (2020), não houve uma predominância na escolha do padrão de metadados, e o DC era

[4] Disponível em: https://ddialliance.org/.
[5] Disponível em: http://schema.datacite.org/.
[6] Disponível em: https://www.openarchives.org/ore/.

adotado por 2 dos 9 repositórios do Brasil, enquanto o DDI era adotado por 1 e o *Ecological Metadata Language* (EML) também por 1.

O DDI é um padrão internacional para descrever os dados produzidos por pesquisas e outros métodos observacionais nas ciências sociais, comportamentais, econômicas e da saúde. Ele é gratuito e pode ser adotado para gerenciar diferentes estágios do ciclo de vida dos dados de pesquisa, como coleta, processamento, distribuição, descoberta e arquivamento. Já o *DataCite* é uma lista das principais propriedades de metadados escolhidas para uma identificação precisa e consistente de um recurso para fins de citação e recuperação, juntamente com instruções de uso recomendadas. Adotado em 6 repositórios, o OAI-ORE define padrões para a descrição e troca de agregações de recursos da Web. Essas agregações, às vezes chamadas de objetos digitais compostos, podem combinar recursos distribuídos com vários tipos de mídia, incluindo texto, imagens, dados e vídeo.

Outros padrões também são adotados nos repositórios do Brasil: 1) o EML, no repositório PPBio *Data Repository*, 2) o *Directory Interchange Format* (DIF), no Redape, e 3) a ISO 19115, voltada para descrição de informações geográficas, no *International Ocean Discovery Program*. Quatro repositórios não indicaram no re3data o padrão de metadados adotado.

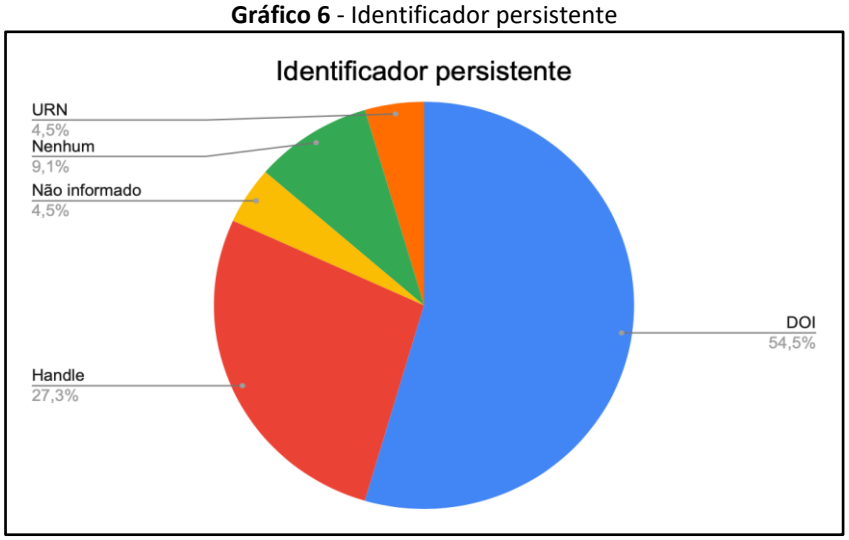
3.7 Identificador persistente

De acordo com Sayão (2007, p. 67), "Um nome persistente, no contexto dos repositórios digitais, é compreendido como um identificador único, o qual deverá perdurar por um período tão longo quanto seja necessário; mesmo que a organização que o atribuiu ao objeto não mais exista quando este for usado". Quando os identificadores estão atrelados à localização ou a processos específicos, há uma instabilidade nos *links,* que leva a mensagens de erro e à incapacidade de acesso ao recurso. Essa instabilidade pode ser contornada com o uso dos identificadores persistentes, que permanecem os mesmos para sempre, independente da localização do recurso (Sayão, 2007).

Como mostra o Gráfico 6, o *Digital Object Identifier* (DOI) é o identificador persistente predominante. Dentro da amostra analisada, 54,5% dos repositórios utilizam o DOI exclusivamente ou em conjunto com outros identificados, como o *Handle*,

presente em 27,3% dos repositórios. Em seguida, observa-se que 9,1% não utilizam nenhum identificador persistente, e 4,5% utilizam o identificador *Uniform Resource Name* (URN), mesmo quantitativo de repositórios que não informaram o identificador persistente que usam.

Gráfico 6 - Identificador persistente



Fonte: Elaborada pelas autoras (2024).
Descrição: Apresenta um gráfico em formato de pizza dividido em cinco partes com cores diferentes em cada uma delas. Tem como título "identificador persistente". A parte maior, representada pelo "DOI", constitui a porcentagem de 54,5%; seguida pelo "handle", com 27,3%; "não informado", com 4,5%; "nenhum", com 9,1%; e URN, com 4,5%.

O identificador DOI foi projetado para uso por máquinas e humanos, criado e regido pela Fundação DOI, que é uma organização sem fins lucrativos, referência em registro de autoridade seguindo a norma ISO 26324 (DOI *Foundation*, 2022). Já o *Handle*, o segundo mais utilizado pelos repositórios brasileiros, também identifica os objetos digitais, a partir dos registros das organizações no Handle.Net[7], sistema administrado pela *Corporation for National Research Initiatives* (CNRI), que é uma organização sem fins lucrativos que busca promover pesquisas de interesse público (Corporation for National Research Initiatives, 2023). O URN pode ser entendido como um conjunto de *Uniform Resource Identifiers* (URI), nomeando uma junção de caracteres para identificar recursos digitais (Sayão, 2007).

Enquanto no estudo de Paganine e Amaro (2020) apenas a metade dos repositórios de dados adotava identificadores persistentes, nesta pesquisa identificou-se que mais de 90% dos repositórios já adotam, com destaque para o DOI. Houve,

[7] Disponível em: https://handle.net/index.html.

portanto, um investimento nesta solução tecnológica que facilita e amplia o acesso contínuo aos dados de pesquisa.

3.8 Versionamento

Dos 19 repositórios da amostra, 16 declararam aceitar versionamento, sendo eles: 1) Deposita Dados, 2) Base de Dados Científicos da Universidade Federal do Paraná, 3) FAPESP COVID-19 *Data Sharing*/BR, 4) PPBio *Data Repository*, 5) Redape, 6) Repositório de Dados de Pesquisa da Unifesp, 7) IBICT *Cariniana Dataverse Network*, 8) *International Ocean Discovery Program*, 9) CEDAP *Research Data Repository*, 10) Repositório Institucional da Unesp, 11) Arca Dados, 12) *FishSounds*, 13) SciELO *Data*, 14) Repositório de Dados de Pesquisa da Unicamp, 15) Repositório de Dados de Pesquisas do Instituto Federal Goiano e 16) Aleia. Isso equivale a 84% dos repositórios analisados.

De acordo com a *World Wide Web Consortium* (W3C), os conjuntos de dados publicados na Web podem mudar ao longo do tempo e, para lidar com essas mudanças, novas versões de um conjunto de dados podem ser criadas no repositório. Em consonância, Rocha *et al.* (2021, p. 14) explicam que "Ao gerenciar versões, o ambiente permite rastrear todas as mudanças que ocorreram no *dataset* ao longo de seu ciclo de vida, isto é, realiza o registro de ações administrativas referentes à proveniência". O ideal é que seja possível atribuir e indicar um número de versão ou data para cada conjunto de dados, de tal forma que humanos e computadores possam facilmente determinar com qual versão de um conjunto de dados eles estão trabalhando.

O software Dataverse possui, por padrão, uma aba denominada "versões", em que registra as alterações feitas no *dataset* e permite que os usuários tenham o controle de qual versão estão acessando. Logo, era esperado que todos os repositórios Dataverse permitissem o versionamento, como é o caso. O resultado encontrado é um ótimo indicativo, já que a indicação de versões melhora a reutilização e a confiabilidade dos *datasets* nos repositórios. Este ponto não foi avaliado no estudo de Paganine e Amaro (2020).

3.9 Política

No Quadro 1, verifica-se que dos 19 repositórios da amostra, 10 possuem algum tipo de política de gestão do repositório. Entende-se como política a definição, seus

objetivos, a indicação dos responsáveis pela implementação e manutenção do repositório e informações sobre o depósito (tipo de material e quem realiza o depósito, por exemplo) (Leite *et al.*, 2013). Um exemplo claro de política pode ser visto no Repositório Institucional da Unesp, que apresenta os responsáveis pela gestão, a missão do repositório, como ele está organizado, os critérios para arquivamento, *etc*.

Alguns repositórios apresentam sua política por meio de portaria ou deliberação, como é o caso do PPBio *Data Repository,* REDU e do Arca Dados. Este último possui uma série de documentos administrativos como o Plano Operativo e a Portaria que institui o repositório. Outros repositórios apresentam-na por meio de um Guias de Usuários, como o caso do Deposita Dados, Redape, SciELO *Data* e Aleia. Apesar desses guias serem voltados para instruir os usuários, eles apresentam os pontos citados por Leite *et al.* (2013), como a organização das comunidades, a equipe responsável pela curadoria, os objetivos do repositório e as informações sobre o depósito.

Quadro 1 - Repositórios que apresentam política de gestão

	Nome do repositório	URL da política
1	GLOBE	http://globe.umbc.edu/documentation-overview/cases-documentation/
2	Deposita Dados	https://depositadados.ibict.br/dvn/guide/guides.html
3	PPBio Data Repository	https://ppbio.inpa.gov.br/sites/default/files/politica_dou.pdf
4	Redape	https://www.redape.dados.embrapa.br/guia-do-usuario.xhtml;jsessionid=09cf9c7327ed6462ddca493c6510
5	Repositório Institucional da UNESP	https://repositorio.unesp.br/server/api/core/bitstreams/95dcc139-3458-4f41-b1f2-5f4a7439878c/content
6	Arca Dados	https://arcadados.fiocruz.br/www/documentos_operativos.html
7	SciELO Data	https://www.scielo.org/en/6.0/user
8	Repositório de Dados de Pesquisa da Unicamp (REDU)	https://www.prp.unicamp.br/wp-content/uploads/sites/4/2022/05/cgpd_delibera_ccp_006_2020_cria_e_regulamenta_o_repositorio_de_dados.pdf
9	Open Research Data @PUC-Rio	https://www.maxwell.vrac.puc-rio.br/projetosEspeciais/ResearchData/WhoWeAre.php?b=2
10	Aleia	https://aleia.ibict.br/about.xhtml

Fonte: Elaborada pelas autoras (2024).

Descrição: Apresenta uma tabela, sob o título “Repositórios que apresentam política de gestão”, constituída por três colunas. A primeira coluna contém os números, de 1 ao 10, correspondente a cada um dos repositórios. A segunda coluna refere-se ao nome dos repositórios, distribuídos também um em cada linha. E a terceira coluna refere-se ao link da política de cada repositório, também distribuído um em cada linha correspondente ao seu repositório.

Outros repositórios como o da Unifesp[8] e o do Instituto Federal Goiano[9] apresentam um conjunto de documentos instrutivos, como guias para criar uma conta, para depositar os dados e para atribuir funções aos usuários. Entretanto, os documentos não apresentam estrutura de política, e não contêm informações administrativas e de gestão como a apresentação do repositório, a equipe responsável, os objetivos, *etc*. Um caso parecido é o da Base de Dados Científicos da Universidade Federal do Paraná[10], que possui um documento de apresentação do repositório, mas em formato de *PowerPoint,* contendo várias definições, a metodologia de implantação da BDC, a linha do tempo de implantação, a seleção de software, a repercussão do repositório, *etc*., não se adequando à descrição de política de gestão trazida por Leite *et al*. (2013). É fundamental que os repositórios de dados de pesquisa tenham a sua política de gestão a fim de documentar todas as informações essenciais, norteando aqueles que acessam e fazem uso do repositório.

4 CONSIDERAÇÕES FINAIS

Neste trabalho o objetivo norteador foi identificar as principais características dos repositórios de dados brasileiros cadastrados no diretório re3data, comparando com o cenário apresentado por Paganini e Amaro (2020). Em comparação com o estudo realizado em 2020, quando o re3data apresentava 9 repositórios indexados, a situação é de crescimento, visto que, no momento, existem 19 repositórios de dados no Brasil, excluindo o Maenduar, que não é um repositório de dados de pesquisa em si, e o *Exploration and Production Data Bank*, o qual não está mais ativo, mas continua mapeado no diretório.

Embora esse quantitativo de repositórios analisados não seja equivalente ao número real que existe no país, é preciso ressaltar o avanço no cenário brasileiro, investindo nessa solução tecnológica. O Brasil é, no momento, o país da América Latina com o maior quantitativo de repositórios de dados indexados no re3data. Na

[8] Disponível em: https://site.unifesp.br/bibliotecas/servicos/guiaperguntasrdp#criar-conta.
[9] Disponível em: https://drive.google.com/file/d/1EOfIZrFzFWw_yVsbrhvHqYNnNJ0hPMMW/view
[10] Disponível em: https://acervodigital.ufpr.br/xmlui/bitstream/handle/1884/56095/Apresenta%c3%a7%c3%a3o_BDC%20%20UFPR.pdf?sequence=1&isAllowed=y

continuidade do estudo, buscar-se-á identificar os repositórios brasileiros existentes e seu cadastrado na plataforma.

Quanto às características levantadas, evidencia-se que há semelhanças com o cenário apresentado por Paganini e Amaro (2020), como o tipo de acesso aos dados: todos os repositórios brasileiros analisados são de acesso aberto. Outra característica que se manteve é a predominância do uso da licença *Creative Commons*, bem como a distribuição homogênea de áreas temáticas.

Em compensação, em quatro anos ocorreram mudanças significativas no contexto dos repositórios de dados do Brasil, como o software adotado: enquanto em 2020 não havia uma predominância na escolha, em 2024 se verifica que o Dataverse é o software mais utilizado, seguido pelo DSpace, justificado principalmente pelo cursos e suporte à infraestrutura oferecida pelo Ibict, pela Rede Nacional de Ensino e Pesquisa (RNP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) com o consórcio CoNCienciA. Outra tendência a ser apontada é o uso do *Dublin Core* junto com outros padrões como o DDI e o *DataCite*, que são específicos para dados de pesquisa. Em 2020, apenas um repositório adotava o DDI e dois o *Dublin Core*.

Em 2020, apenas metade dos repositórios adotava um identificador persistente. Já em 2024, 16 dos 19 repositórios analisados apresentaram um, sendo o DOI o mais utilizado, seguido do *Handle*. Destaca-se também que o próprio re3data está passando por revisões, e muitos dos repositórios podem ter alterado suas formas de atribuição de identificador persistente sem a alteração na base. Quanto ao versionamento, o resultado é semelhante: 16 dos 19 repositórios permitem o versionamento, uma característica que não foi apresentada no estudo de 2020.

Essas mudanças apontam para o crescimento e aprimoração dos repositórios de dados no contexto da Ciência Aberta, promovendo o acesso aberto ao conhecimento científico, que envolve não só os resultados das pesquisas, mas também outros insumos valiosos como os próprios dados coletados. A existência de repositórios de dados em acesso aberto diminui barreiras para os pesquisadores e consequentemente para toda a sociedade, de modo a atenuar os impactos de obstáculos geográficos e financeiros. Por isso a importância de mapeá-los e identificar suas principais características, auxiliando a traçar um perfil das soluções adotadas no Brasil.

Ressalta-se que esta pesquisa limitou-se aos repositórios de dados indexados no re3data, bem como aos dados declarados no diretório pelos repositórios. Sendo assim, não representa a totalidade do cenário do Brasil, mas aponta as tendências dos repositórios que estão mapeados internacionalmente como um representativo do país.

REFERÊNCIAS

CABALLERO-RIVERO, A.; SANCHEZ-TARRAGO, N.; SANTOS, R. N. M. Práticas de Ciência Aberta da comunidade acadêmica brasileira: estudo a partir da produção científica. **Transinformação**, v. 31, e190029, 2019. Disponível em: https://www.scielo.br/j/tinf/a/5hgYK97mbcjRdZL7dfRDzvD/?lang=pt#. Acesso em: 24 jun. 2024.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES (CNRI). **Handle.NetRegistry**, 2023. Serviço de informação. Disponível em: https://handle.net/index.html. Acesso em: 27 jun. 2024.

DATAVERSE. **The Dataverse Project**. [*S.l.*], 2020. Disponível em: https://dataverse.org. Acesso em: 24 jun. 2024.

DOI FOUNDATION. **DOI**, 2022. The Foudation about us. Disponível em: https://www.doi.org/the-foundation/about-us/. Acesso em: 27 jun. 2024.

EUROPEAN COMMISSION. **Guidelines on FAIR data management in horizon 2020**. 2016. Disponível em: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Acesso em: 8 abr. 2024.

LEITE, Fernando *et al*. **Boas práticas para a construção de repositórios institucionais da produção científica**. Brasília, DF: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2013. Disponível em: https://livroaberto.ibict.br/bitstream/1/703/1/Boas%20pr%c3%a1ticas%20para%20a%20constru%c3%a7%c3%a3o%20de%20reposit%c3%b3rios%20institucionais%20da%20produ%c3%a7%c3%a3o%20cient%c3%adfica.pdf. Acesso em: 26 jun. 2024.

PAGANINE, Lucas Nóbrega; AMARO, Bianca. Características dos repositórios de dados científicos no Brasil. **Biblos**, v. 34, n. 1, p. 176-188, dez. 2020. Disponível em: https://doi.org/10.14295/biblos.v34i1.11132. Acesso em: 21 jun. 2024.

RAMACHANDRAN, Rahul; BUGBEE, Kaylin; MURPHY, Kevin. From open data to open science. **Earth and Space Science**, v. 8, n. 5, p. e2020EA001562, 2021. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2020EA001562. Acesso em: 10 jul. 2024.

REIS, Juliani Menezes dos; ROZADOS, Helen Beatriz Frota. O livro digital o direito autoral à luz do Copyleft, Creative Commons e Digital Right Management. **Biblos**:

Revista do Instituto de Ciências Humanas e da Informação, v. 27, n. 2, p. 63-77, jul./dez. 2013. Disponível em: https://lume.ufrgs.br/handle/10183/183682. Acesso em: 26 jun. 2024.

ROCHA, Rafael Port *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://brapci.inf.br/#/v/160963. Acesso em 27 jun. 2024.

RODRIGUES, Marcello Mundim; DIAS, Guilherme Ataíde; LOURENÇO, Cíntia de Azevedo. Repositórios de dados científicos na américa do sul: uma análise da conformidade com os princípios fair. **Em Questão**, Porto Alegre, v. 28, n. 2, p. 113057, 2022. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/113057. Acesso em: 8 jul. 2024.

RÜCKNAGEL, Jessika *et al*. **Metadata Schema for the Description of Research Data Repositories: version 3.1**. Re3data, 2021. Disponível em: https://gfzpublic.gfz-potsdam.de/rest/items/item_1397899_10/component/file_1398549/content. Acesso em: 25 jun. 2024.

SAGIROGLU, Seref; SINANC, Duygu. Big data: A review. **Conferência Internacional sobre Tecnologias e Sistemas de Colaboração (CTS) de 2013**, San Diego, CA, EUA, 2013, p. 42-47. Disponível em: https://ieeexplore.ieee.org/document/6567202. Acesso em: 21 jun. 2024.

SANCHEZ, Fernanda Alves; VIDOTTI, Silvana Aparecida Borsetti Gregório; VECHIATO, Fernando Luiz. A contribuição da curadoria digital em repositórios digitais. **Revista Informação na Sociedade Contemporânea**, [*s.l.*], v. 1, p. 1-17, 2017. Disponível em: https://periodicos.ufrn.br/informacao/article/view/12280. Acesso em: 21 jun. 2024.

SAYÃO, Luís Fernando. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes -URN, PURL, DOI, Handle System, CrossRef e OpenURL. **TransInformação**, Campinas, v. 19, n.1, p. 65-82, jan./abr., 2007. Disponível em: https://www.scielo.br/j/tinf/a/NTr5XbPG7LG5pWH876MmWVN/?format=pdf. Acesso em: 27 jun. 2024.

SAYÃO, Luis Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90-115, maio/ago., 2016. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27939/20122. Acesso em: 20 jun. 2024.

SOUZA, Marcia Izabel Fugisawa; VENDRUSCULO, Laurimar Gonçalves; MELO, Geane Cristina. Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core. **Ciência da Informação**, Brasília, DF, v. 29, n. 1, p. 93-102, jan./abr. 2000. Disponível em: https://www.scielo.br/j/ci/a/tcW3q4WvNBQNTqTyLK8qfFF/#:~:text=Dublin%20Core%20pode%20ser%20definido,a%20descri%C3%A7%C3%A3o%20de%20recursos%20eletr%C3%B4nicos.. Acesso em: 27 jun. 2024.