ISSN: 2237-826X
ARTIGO | PAPER

Produção científica sobre repositório de dados de pesquisa: representações autorais e temáticas

Overview of scientific production on data repository: copyright and thematic representations

Igor Yure Ramos Matos[1], Divino Ignacio Ribeiro Júnior[2], Jorge Kroll do Prado[3], Julibio David Argino[4]

[1] *Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil. ORCID: https://orcid.org/0000-0001-7230-5100*

[2] *Universidade do Estado de Santa Catarina (UDESC), Florianópolis, Santa Catarina, Brasil. ORCID: https://orcid.org/0000-0002-9705-6507*

[2] *Universidade do Estado de Santa Catarina (UDESC), Florianópolis, Santa Catarina, Brasil. ORCID: https://orcid.org/0000-0002-7287-8133*

[2] *Universidade do Estado de Santa Catarina (UDESC), Florianópolis, Santa Catarina, Brasil. ORCID: https://orcid.org/0000-0002-9114-6229*

Autor para correspondência/*Mail to:* Igor Yure Ramos Matos, igoryure.rm@gmail.com

Recebido/*Submitted*: 18 de março de 2022; **Aceito/*Approved*:** 20 de outubro de 2022

Resumo

Introdução: trata-se de um levantamento da produção científica na Ciência da Informação em torno dos repositórios de dados de pesquisa. A produção científica é fundamental para entender o estado da arte e o seu desenvolvimento ao longo dos anos, principalmente de temas emergentes como aqui propostos. Os dados de pesquisa são os insumos principais para o fazer científico na *e-Science*, preocupada com o reuso, compartilhamento, colaboração, economia, intercâmbio e rapidez das pesquisas. **Objetivo:** apresentar um panorama da produção científica na Ciência da Informação em torno do tema repositórios de dados de pesquisa. **Método:** esta é uma pesquisa exploratória, bibliográfica e quantitativa a partir dos resultados encontrados nas seguintes fontes de informação: BRAPCI, SciELO, Portal de Periódicos da CAPES, BDTD e Lisa. **Resultados:** recuperou-se um total de 127 trabalhos publicados em periódicos científicos, com 290 autores diferentes. A literatura em inglês começou em 2009 a publicar relatos de repositório de dados, sendo recuperados trabalhos sobre metadados e padrões de interoperabilidade de dados desde 2005. Percebe-se que o primeiro artigo brasileiro recuperado nas buscas é de 2015. Já 2019 constitui-se como o ano em que mais se publicou sobre o tema, com nove artigos, contra cinco publicados em inglês.**Conclusão:**: é um tema emergente e com produção crescente dentro da CI no Brasil, necessitando de bibliotecários com conhecimentos e habilidades para a gestão de repositórios e dados de pesquisa.

Palavras-chave: Repositório de dados; Gestão de dados de pesquisa; Ciência da Informação; Ciência Aberta.

Abstract

Introduction: *this is a survey of scientific production in Information Science around research data repositories. Scientific production is essential to understand the state of the art and its development over the years, especially in emerging themes as proposed here. Research data are the main inputs for scientific work in e-Science, concerned with the reuse, sharing, collaboration, economy, exchange and speed of research.* ***Objective:*** *to present an overview of scientific production in Information Science around the topic of research data repositories.* ***Method:*** *this is an exploratory, bibliographical and quantitative research based on the results found in the following sources of information: BRAPCI, SciELO, Portal de Periódicos da CAPES, BDTD and Lisa.****Results:*** *a total of 127 works published in scientific journals were retrieved, with 290 different authors. The literature in English began in 2009 to publish data repository reports, and works on metadata and data interoperability standards have been retrieved since 2005. It can be seen that the first Brazilian article retrieved in the searches is from 2015. the year that published the most on the subject, with nine articles, against five published in English.* ***Conclusions:****: it is an emerging topic with increasing production within the CI in Brazil, requiring librarians with knowledge and skills to manage repositories and research data.*

Keywords: *Data repositories; Management Research data; Information Science; Open access.*

INTRODUÇÃO

O fazer científico é pensado no trabalho colaborativo, no reuso e no compartilhamento dos dados de pesquisa para a construção do conhecimento científico. Os dados publicados em periódicos, livros, teses, dissertações e demais publicações científicas são somente a ponta do iceberg, pois grande parte fica submersa ou perdida (Rocha & Schmidt, 2011). A disponibilidade integral dos dados permite que outros pesquisadores cheguem a outras conclusões e descobertas que o coletor/autor original desses não visualizou/descobriu. Para Henning, Ribeiro, Sales, Moreira, e Santos (2019, p. 176), "as novas formas de fazer ciência, pautadas fortemente no compartilhamento e no reuso de dados de pesquisa, vêm colocando em evidência a necessidade de deixar para trás a ideia dos dados apenas como insumos intermediários das atividades científicas."

Essa mudança científica é conhecida por *e-Science* e se caracteriza pela intensa coleta de dados, utilizando instrumentos, sensores e computadores. Em palestra de Gray transcrita por Rocha e Schmidt (2011), o *e-Science* é o quarto paradigma da ciência, devido à grandeza e infinidade de dados coletados diariamente, com enorme valor e poder para o desenvolvimento científico, econômico e industrial. O compartilhamento e reuso desses são uma das práticas do *e-Science*, cujas vantagens são: a) economia de tempo dos pesquisadores; b) economia

dos gastos com instrumentos e equipamentos de laboratórios, como: computadores, câmeras, sensores, drones, microscópios, lupas, aparelhos de ressonância e inúmeros aparelhos com alto valor comercial; c) evita-se a duplicação da coleta de dados; e, d) pesquisadores são citados e reconhecidos quando seus dados são reutilizados (Rocha & Schmidt, 2011; Semeler, 2019; Silva, 2019).

Os dados de pesquisa são todos os elementos coletados para subsidiar e comprovar hipóteses de pesquisa ou atingir os objetivos propostos. Estes apoiam as publicações primárias, ou seja, os elementos básicos que dão veracidade e comprovam a revisão de literatura, alvejando os objetivos da pesquisa. Segundo Silva (2019, p. 3), os dados são valiosos porque "[...] se o conhecimento é o motor do avanço científico, os dados são seu combustível." Para que o compartilhamento e reuso tornem-se realidade, surgem os repositórios de dados como infraestruturas tecnológicas para o depósito de dados de pesquisa. Segundo Sanchez, Vechiato, e Vidotti (2019, p. 52), "os repositórios de dados [são] ambientes informacionais digitais que buscam armazenar, organizar, representar, prover acesso, disseminar e preservar dados oriundos de pesquisas científicas."

> A gestão dos repositórios de dados de pesquisa deve ficar a cargo dos "bibliotecários de dados (Semeler, 2019)." São bibliotecários que se especializarão em realizar a curadoria dos dados de pesquisa. Os repositórios de dados e a gestão de dados são novas demandas para as bibliotecas universitárias, segundo Oliveira e Silva (2016, p. 6) "[...] a realidade brasileira voltada para a ciência aberta e dados de pesquisa encontra-se em um estágio incipiente", necessitando assim de novas publicações e profissionais interessados em adquirir conhecimentos e habilidades no tema.

Nesta perspectiva, este artigo tem por objetivo apresentar um panorama da produção científica na Ciência da Informação em torno do tema repositórios de dados de pesquisa. Seus objetivos específicos contemplam: (a) conhecer quais os autores que estão publicando sobre gestão de dados e repositórios de dados; (b) identificar os periódicos que mais publicaram o tema; (c) conhecer os assuntos importantes a partir das palavras-chave empregadas nos artigos recuperados.

REVISÃO DE LITERATURA

Os dados de pesquisa são elementos coletados para subsidiar as pesquisas, confirmar ou refutar hipóteses, são os insumos que a ciência utiliza para novas descobertas, por esse motivo, os dados são tão importantes. Para Sayão e Sales (2016, p. 90-91), os dados deixam de ser simples subprodutos das atividades de pesquisa e se tornam recursos informacionais de primeira grandeza, caracterizando um novo paradigma científico pautado pelo compartilhamento, amplo acesso e reuso de dados. Pavão et al. (2019, p.7) definem os dados de pesquisa como "[...] dados coletados, observados ou produzidos durante a pesquisa (números, textos, imagem, som, saídas de equipamentos) para fins de análise e produção de resultados de pesquisa originais." Existe uma infinidade de tipos de dados de pesquisa, porque são elementos presentes em todas as áreas acadêmicas. Sua coleta pode tanto utilizar de tecnologias, como *softwares* e demais instrumentos tecnológicos, já em formato digital, quanto coletados manualmente (depois tratados e digitalizados) por inúmeros cientistas e armazenadas em um único *database*, por isso, é uma prática cooperativa e colaborativa.

> Dados de pesquisa são os materiais comumente registrados e aceitos na comunidade científica como necessários para validar os resultados de pesquisa e incluem: fatos e estatísticas recolhidas para posterior referência ou análise, documentos (texto, Word), planilhas (Excel, etc.), cadernos de laboratório, cadernos de campo, diários, questionários, transcrições, fitas de áudio, fitas de vídeo, fotografias, filmes, sequências de proteínas ou genéticos, respostas de teste, slides, artefatos, amostras, coleção de objetos digitais adquiridos e gerados durante o processo de pesquisa, conteúdos de banco de dados (vídeo, áudio, texto, imagens), modelos, algoritmos, *scripts*, arquivos de *log*, *software* de simulação, metodologias e fluxos de trabalho, procedimentos operacionais, padrões e protocolos. (Dudziak, 2016, não paginado).

Os dados de pesquisa devem atender aos princípios FAIR – *Findable* (encontrável), *Accessible* (acessível), *Interoperable* (interoperável) e *Reusable* (reutilizável). São as características que os dados precisam para adquirir qualidade com o objetivo do reuso. Ser "Encontrável" diz respeito a passíveis de serem conhecidos/recuperáveis/encontráveis por outros pesquisadores, assim, necessita da correta descrição dos seus metadados. Ser "Acessível" relaciona-se com estar depositados em um repositório de dados, principalmente em acesso aberto e que utilizem de identificadores persistentes para preservação e acessíveis a longo prazo. Ser "Interoperável" relaciona-se com a capacidade de reuso também por outros computadores/*softwares*, sem necessariamente a interferência humana, ou seja, que cumpra o protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). Ser "Reutilizável" diz respeito a estar em formato aberto, possuir documentação (o Plano de Gestão de Dados (PGD) e um glossário de termos, por exemplo) para serem entendidos pelos pesquisadores e/ou por outros *softwares*. Os princípios FAIR são recomendações de boas práticas para possibilitar que os dados sejam compartilhados com qualidade, consistência e integridade.

A gestão de dados de pesquisas constitui-se de procedimentos e ações necessárias para que os dados tenham consistência, integridade e qualidade, com a finalidade do compartilhamento e reuso. São passos que envolvem

desde o planejamento da pesquisa, a execução/coleta, o formato, os *softwares* de análise e limpeza, a descrição dos metadados, o armazenamento, até o depósito e preservação dos dados. Segundo Silva (2019, p. 3), "[...] para os pesquisadores, uma gestão adequada dos dados de pesquisa permite novas maneiras de comparação e de descobrimentos, isto é, permite gerar novos campos de pesquisa."

> A gestão de dados científicos[1] é um conjunto de atividades que visa a coletar, armazenar, gerenciar e compartilhar dados provenientes de pesquisa científica. Uma gestão de dados eficaz possibilita a racionalização de recursos, por meio do reuso e compartilhamento de dados. Na USP, a gestão de dados científicos tem a finalidade de auxiliar o pesquisador em relação a: planejamento, organização e segurança; documentação e compartilhamento; preparação dos conjuntos de dados para depósito; preservação dos dados; questões relacionadas a direitos autorais; licenciamento e propriedade intelectual. (Universidade de São Paulo, n.d., não paginado).

Para isso, os pesquisadores que coletam dados necessitam adotar algum modelo de ciclo de vida de dados, em que estão as etapas para garantir a qualidade do conjunto de dados e a documentação adequada. Os modelos de ciclo de vida de dados são modelos teóricos, alguns gerais, aplicados a todas as áreas do conhecimento, e outros específicos. Um exemplo de modelo geral é o DATAONE, que é cíclico e cujas etapas contêm várias ações e práticas.

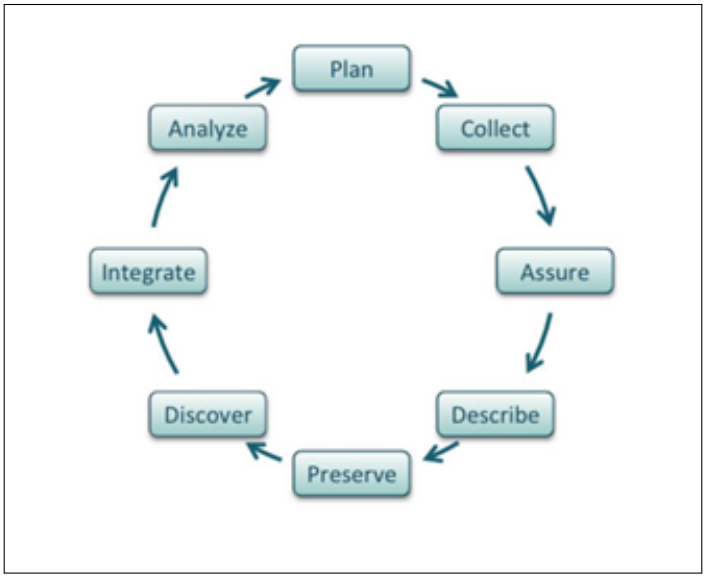

Figura 1. Ciclo de dados de pesquisa DATAONE.
Fonte: Strasser, Cook, Michener, e Budden (2012, p. 3).

O ciclo de vida dos dados é um dos principais componentes da gestão de dados, no que se relaciona às fases de coleta até a preservação dos dados. Outros modelos são: *Curation Lifecycle Model*[2], *DDI Lifecycle Model*[3] e o *Research Data Lifecycle*[4]. Outro importante componente da gestão é o Plano de Gestão de Dados (PGD), um documento elaborado na fase inicial do planejamento da pesquisa. Ele deve anteceder a coleta de dados, descrever a metodologia, objetivos, toda a coleta, tipos e formato de dados, instrumentos, data da coleta, tipos de *software* e versão utilizada, ou seja, todas as informações importantes para que o conjunto de dados faça sentido a outros pesquisadores. É um documento realizado pelos pesquisadores e deve ser atualizado durante o projeto e disponibilizado junto com o conjunto de dados nos repositórios de dados.

> O PGD descreve o ciclo de vida de gestão para todos os dados que serão coletados, processados ou gerados por um projeto de pesquisa. De uma forma abreviada, ele se constitui em um documento formal que estabelece um compromisso de como esses dados serão tratados durante todo o desenvolvimento do projeto, e também após a sua conclusão. Para isso, o PGD descreve, de uma forma geral, que dados serão processados, coletados ou gerados; quais as metodologias e padrões que serão utilizados nesses processos; se, como e sob que condições esses dados serão compartilhados e/ou tornados abertos para a comunidade de pesquisa; e como eles serão curados e preservados. (Sayão & Sales, 2015, p. 15).

Para a realização do PGD, tem-se duas plataformas/ferramentas *online* que permitem a confecção deste documento: *DMPTool* (https://dmptool.org/) e *DMPonline* (https://dmponline.dcc.ac.uk/). Ambas possuem

[1] O termo adotado neste trabalho foi dados de pesquisa, existe uma divergência em relação alguns autores.
[2] The Curation Lifecycle Model. Disponível em: https://www.dcc.ac.uk/guidance/curation-lifecycle-model.
[3] DDI (3.3) Documentation. Disponível em: https://ddi-lifecycle-documentation.readthedocs.io/en/latest/User%20Guide/Introduction.html
[4] Research Data lifecycle. Disponível em: https://ukdataservice.ac.uk/learning-hub/research-data-management/. Acesso: 15 jul. 2022.

a mesma estrutura de perguntas para a descrição dos planos, divididas em: coleta de dados, documentação e metadados, ética e conformidade legal, armazenamento e backup, seleção e preservação, compartilhamento de dados e responsabilidade e recursos.

Os dados de pesquisas são protegidos pelos direitos de propriedade intelectual, especificamente os direitos autorais. Na ciência aberta recomenda-se a aplicação de licenças abertas cujo uso o próprio autor permite, compartilhamento, modificações ou uso irrestrito. Dois tipos de licenças abertas são as *Creative Commons* e *Open Data Commons*. A primeira possui diferentes tipos de permissões, que concedem o direito de uso, somente necessitando dar os devidos créditos a seus autores e não infringir a finalidade que a licença permite. A segunda é um contrato de licença destinado a permitir que os usuários compartilhem, modifiquem e usem livremente para qualquer finalidade e sem quaisquer restrições, sendo específica para a reutilização de bancos e conjuntos de dados.

Na parte tecnológica, a disponibilização dos dados de pesquisa ocorre em plataformas de *softwares* para repositórios de dados. Segundo Moreno (2018, p. 53), "[...] a criação de infraestrutura e manutenção de repositórios de dados de pesquisa está em curso em diversos países e apresenta-se como um desafio tanto em termos de gestão quanto da representação dos dados que estão contidos nesses sistemas." Diversas são as tecnologias disponíveis para a construção de repositórios de dados.

> Os repositórios de dados são mantidos por conjuntos de ações que viabilizam o armazenamento de dados visando à otimização da recuperação, o que amplia as potencialidades de reuso destes dados entre os pesquisadores. Desta forma, agiliza os processos de investigação e, consequentemente, o avanço na ciência. Com uma infraestrutura implementada por repositórios de dados, apoiada por um Plano de Gerenciamento de Dados (PGD) bem fundamentado, os pesquisadores têm aporte propício para depósito de seus conjuntos de dados e busca e recuperação de dados já coletados por outros pesquisadores, que poderão ser reutilizados em suas pesquisas. (Monteiro, 2017, p. 15)

Segundo Semeler (2019, p. 138), "os repositórios de dados de pesquisa são o locus de ligação entre os usuários de dados de pesquisa (pesquisadores) e os bibliotecários de dados. Neles, os bibliotecários de dados podem oferecer serviços e produtos relacionados ao acesso e à preservação de dados de pesquisa." Segundo Pavão, Rocha, e Gabriel Junior (2018, p. 331), o Brasil ainda não possui uma estrutura de apoio à construção de repositório de dados e nem um planejamento a nível nacional para esse apoio.

PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa caracteriza-se como bibliográfica e exploratória. A pesquisa bibliográfica se caracteriza por pesquisar em documentos já publicados, segundo Prodanov e Freitas (2013, p. 54), "constituído principalmente de: livros, revistas, publicações em periódicos e artigos científicos, jornais, boletins, monografias, dissertações, teses, material cartográfico, internet, com o objetivo de colocar o pesquisador em contato direto com todo material já escrito sobre o assunto da pesquisa". E por pesquisa exploratória, segundo Severino (2017, [p. 68]), "busca apenas levantar informações sobre um determinado objeto, delimitando assim um campo de trabalho, mapeando as condições de manifestação desse objeto." Para Cervo e Bervian (2002, p. 69), os estudos exploratórios familiarizam-se com o fenômeno ou obter nova percepção do mesmo e descobrir novas ideias.

Definição de Termos

Para o começo das buscas, confeccionou-se uma lista dos termos (Tabela 1) que mais apareciam em artigos conhecidos sobre o tema, permitiu-se assim identificar o assunto principal e os secundários[5] relacionados com os repositórios de dados. Sendo os dados de pesquisa um tema incipiente (Oliveira & Silva, 2016), logo também o são os repositórios de dados. Assim, como critério, os artigos selecionados são publicações que abrangem todo o conhecimento necessário para implantação de um repositório de dados, por exemplo, instalação do *software*, a gestão de dados, padrões de metadados e identificadores persistentes, etc. A lista foi traduzida nos três idiomas com os quais os autores possuem mais afinidade.

[5] Assunto principal refere-se a experiências de implantação de *software* e customizações. Assunto secundário é considerado o tema que não tem relação direta com o objetivo principal da pesquisa do mestrado, mas importantes para se construir e gerenciar um repositório de dados.

Português	Inglês	Espanhol
Repositório de dados	*Data repository*	*Repositorios de datos*
Dados de pesquisa	*Research data*	*Datos de investigación*
Ciência aberta	*Open Science*	*Ciencia abierta*
Curadoria digital	*Data curation*	*Curaduría de datos*
Gestão de dados de pesquisa	*Data management*	*Gestión de datos de investigación*
Propriedade intelectual	*Intellectual property*	*Propriedad intelectual*
Plano de Gerenciamento de dados	*Data management plan*	*Plan de gestión de datos*
Licenças de uso	*Use license*	*Licencia de uso*
Direito autoral	*Copyright*	*Derecho de autor*

Tabela 1. Termos para a pesquisa.

A composição da tabela 1 contribuiu na seleção dos artigos nas bases de dados pesquisadas, porém poucos dos termos foram utilizados nas estratégias de busca. Torna-se válido mantê-la para que iniciantes no tema percebam os temas correlatos aos repositórios de dados.

Estratégia de Busca

No início do levantamento em base de dados, pesquisou-se em quatro bases de periódicos ou de resumos de periódicos e um repositório de teses e dissertações, sendo: a) Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI); b) *Scientific Electronic Library Online* (SciELO); c) Biblioteca Digital de Teses e Dissertações (BDTD); d) *Library and Information Science Abstracts* (LISA); e) Portal de periódicos CAPES.

A busca na BDTD recuperou-se três teses e duas dissertações, já na LISA, nove trabalhos de teses e dissertações. Mesmo tendo realizado busca na BDTD, decidiu-se posteriormente que neste analisaria somente artigos, deixando teses e dissertações para estudos futuros. Um dos fatores decisivos foram poucos trabalhos recuperados, ao ponto que os artigos estão em maior quantidade.

a Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI): A busca na BRAPCI foi realizada em 03 de janeiro de 2021, com estratégia de busca "Repositório de dados AND Ciência da Informação", com delimitação da busca de 1972 a 2021, com ordenação por relevância, sendo recuperados 36 documentos. Foi exportado em formato de documento (.DOC) e *Extensible Stylesheet Language* (XSL). Consultou-se o tesauro da base para entender as estratégias de pesquisa. Foram excluídos dossiês editoriais e um trabalho duplicado. A BRAPCI apresentou alguns erros na planilha, como: apresentar após o título "@pt-BR" nas palavras-chave a base acrescenta termos como: "Ciência da Computação", "Biblioteconomia", "Ciência da Informação", assim as palavras-chave recuperadas não são somente as que o autor utilizou no artigo.

b *Scientific Electronic Library Online* (SciELO): A busca na SciELO foi realizada em 05 de janeiro de 2021. Tentou-se algumas estratégias de busca, sendo a que mais retornou itens foi o termo: "repositório de dados", com 49 itens e, após analisados, somente cinco foram exportados. Realizou-se outra busca em 11 de janeiro, utilizando-se na pesquisa simples em todos os índices: "repositório de dados AND ciência da informação", com oito trabalhos (Figura 2), sendo sete selecionados e exportados para CSV e três repetidos com a primeira busca. O resultado dos dois dias de busca totalizou nove artigos recuperados na SciELO e, após leitura técnica, somente quatro artigos permaneceram.

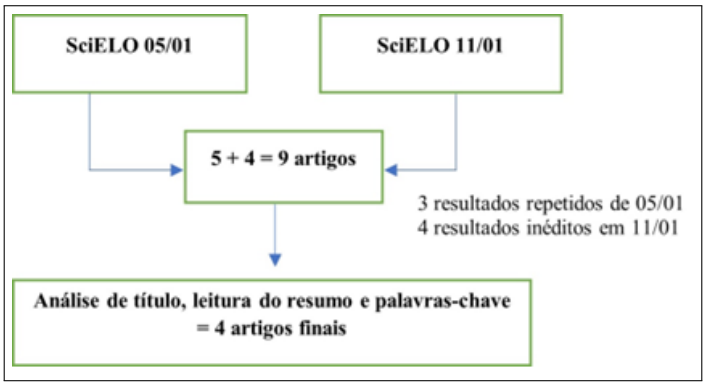

Figura 2. Etapas da seleção dos resultados da SciELO.

c Biblioteca Digital de Teses e Dissertações (BDTD): A busca avançada na BDTD foi realizada em 11 de janeiro de 2021, com os termos: "repositório de dados e ciência da informação", utilizando todos os campos

e correspondência da busca: todos os termos, recuperaram-se "6.795 itens". Nova busca utilizando-se os mesmos termos, delimitando por assunto, a busca não recuperou nenhum trabalho. Termos de busca: "(Assunto: repositório de dados E Assunto: ciência da informação)". Delimitando por título, também não foi recuperado nenhum trabalho. Em outra tentativa de busca (Título: repositório de dados E Título: ciência da informação) – não recuperou nenhum registro. Outra tentativa foi realizada com os mesmos termos: repositório de dados e ciência da informação, delimitando o primeiro termo por título e o segundo termo por assunto, recuperou-se um trabalho. Nova tentativa realizada na BDTD na mesma data utilizando de busca simples, com os termos entre aspas e o operador booleano AND, "repositório de dados" AND "ciência da informação", recuperou seis trabalhos, sendo uma tese duplicada, portanto três teses e duas dissertações. Foi exportado em formato *Comma-Separated-Values* (CSV) para análise posterior, porém em uma leitura mais rápida, somente quatro trabalhos têm relevância para esta pesquisa.

d *Library and Information Science Abstracts* (LISA): A busca foi realizada em 21 de janeiro de 2021, com os termos "*Data repository* AND "*Information Science*", recuperando 10.341 resultados. Percebeu-se que recuperou trabalhos muitos genéricos e com baixo grau de precisão. Desta forma, pesquisou-se com a seguinte estratégia: "*Data Repository*" AND *Information Science*, recuperando 444 resultados. Analisando os resultados, alguns traziam "informação" e "ciência" como assunto, e não "Ciência da Informação" em si. Refez-se a busca utilizando "*Data repository*" AND "*Information Science*", recuperando-se 215 resultados. Após leitura técnica de título, resumo e palavras-chave, permaneceram 94 trabalhos.

e Portal de periódicos CAPES: Realizou-se o *login* com a rede VPN da UFSC para acesso ao portal em 12 de janeiro de 2021 com os "*Data Repository*" AND "*Information Science*". Recuperaram-se 54 artigos que, após primeira leitura de título e resumo, exportaram 27 registros. Realizou-se uma leitura técnica mais detalhada com título, resumos e palavras-chave, permanecendo 15 trabalhos ao final.

Tratativa dos Resultados

Os dados foram coletados e salvos em formato de planilha ou documento. Em uma nova planilha, realizou-se a limpeza e organização dos metadados, mantiveram-se os seguintes campos, cada qual corresponde a uma coluna: título do artigo em português e inglês, autor(es), título do periódico, volume e número do periódico, *International Standard Serial Number* (ISSN), *Digital Object Identifier* (DOI) e *Uniform Resource Locator* (URL), ano, palavras-chave em português e em inglês e nome da base. Alguns dos metadados estavam incorretos, assim, realizou-se a conferência manual, corrigindo, confirmando ou capturando-os novamente. Os artigos em inglês que não tinham palavras-chave em português foram traduzidas utilizando a ferramenta tradutora do Google.

Base	Data da busca	Estratégia	Trabalhos exportados	Planilha final
BRAPCI	03/01/2021	"Repositório de dados AND Ciência da Informação"	36 resultados	23 documentos no final, 7 repetidos nos periódicos CAPES, 1 repetido da Lisa
SciELO	05/01/2021 e 11/01/2021	repositório de dados AND ciência da informação	5 resultados 8 resultados	Somente 04 trabalhos no final
BDTB	11/01/2021	"repositório de dados" AND "ciência da Informação"	6 resultados	Optou-se por não utilizar teses e dissertações neste trabalho.
Portal de Periódicos CAPES	12/01/2021	"*Data repository*" AND "*Information Science*"	15 resultados	15 documentos no final, 7 repetidos com a Lisa, 1 repetido com a BRAPCI
LISA	21/01/2021	"*Data repository*" AND "*Information Science*"	215 resultados	94 artigos no final, 7 repetidos com a CAPES, 1 repetido entre CAPES e BRAPCI

Tabela 2. Visão geral das bases e estratégias.

RESULTADOS

Obtiveram-se na planilha final 127 artigos[6] recuperados nas buscas nas bases da BRAPCI, Portal de periódicos CAPES, SciELO e LISA. Dividiu-se a análise dos resultados em quatro grandes blocos: a) autores; b) títulos de periódicos; c) artigos: idioma e ano, e d) palavras-chave.

[6]Dentre esses, veio uma entrevista publicada em um periódico brasileiro, com dois profissionais que trabalham com repositório de dados e/ou gestão de dados de pesquisa. E um outro trabalho publicado nos anais do ENANCIB. Como eram trabalhos relevantes e publicados com revisão dos pares, mantiveram-se na análise.

A escolha em analisar por estes blocos ocorreu por entender que estes são os dados mais importantes para caracterizar os estudos sobre o tema.

Autores

Em relação aos autores, recuperaram-se 290 diferentes, sendo que 19 autores com dois ou mais artigos publicados (Tabela 3) sobre o tema de repositório de dados ou temas secundários, como: curadoria de dados, gestão de dados de pesquisa, padrões de metadados, ciência aberta, entre outros. Percebe-se que cinco autores são os mais produtivos sobre o tema e estão em três países: Coreia do Sul, Brasil e EUA. Observa-se que três autores são autores/coautores de 11 artigos (8,7%) dos 127 artigos recuperados, sendo que 19 autores (6,6%) são os mais produtivos sobre o tema, com 34 trabalhos que correspondem a 26,7% dos trabalhos recuperados.

Autor	País	N	% Autores	Artigos	Artigos Acumulados
Luana Farias Sales	BR	1	0,3%	6	6
Luís Fernando Sayão[a]	BR	2	0,7%	5	6
Youngseek Kim	KOR	3	1,0%	5	11
Lisa R. Johnston	EUA	4	1,4%	3	14
Suntae Kim	KOR	5	1,7%	3	17
Abigail Goben	EUA	6	2,1%	2	19
Caterina Marta Griposo Pavão[b]	BR	7	2,4%	2	21
Elizabete Cristina de S. de A. Monteiro[c]	KOR	8	2,8%	2	23
Jane Cho	KOR	9	3,1%	2	25
Jeremy Kenyon	EUA	10	3,4%	2	27
Li Si	CHN	11	3,8%	2	29
Lin He	CHN	12	4,1%	2	31
Rafael Port da Rocha	BR	13	4,5%	2	31
Rene Faustino Gabriel Junior	BR	14	4,8%	2	31
Ricardo Cesar Gonçalves Sant'Ana	BR	15	5,2%	2	31
Sônia Elisa Caregnato	BR	16	5,5%	2	32
Wei Jeng	TWN	17	5,9%	2	34
Wenning Xing	CHN	18	6,2%	2	34
Xiaohe Zhuang	CHN	19	6,6%	2	34

Tabela 3. Autores que mais publicaram sobre o tema.

[a]A coluna de artigos acumulados referentes a esse autor não foi quantificado, pois escreveu em coautoria com Sales.
[b]A autora escreveu em coautoria com Rafael Port e Rene Faustino.
[c]Os dois artigos foram escritos em coautoria com Ricardo Cesar Gonçalves Sant'Ana.

Pode-se apresentar de outra maneira os autores (Figura 3): teve-se um total de 321 em 127 artigos. Identificaram-se 290 autores diferentes, sendo que destes, 271 escreveram um único artigo, 15 escreveram dois artigos, dois escreveram cinco e um único autor foi o mais produtivo, com seis artigos.

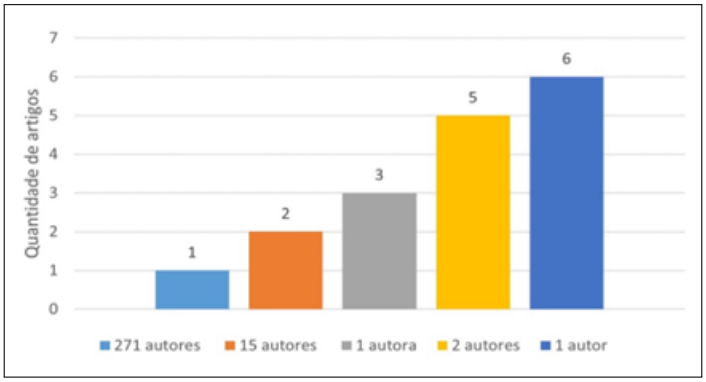

Figura 3. Produtividade dos autores no tema.

Em relação à quantidade de autores por artigo (Figura 4), verificou-se que a média é de 44 artigos escritos por dois autores. Percebe-se que na área publica-se com até quatro autores, raramente ultrapassando esse número.

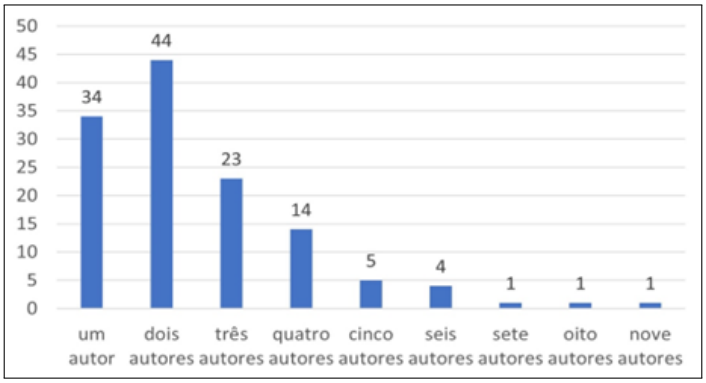

Figura 4. Número de autores por artigo.

Os artigos foram recuperados em bases da Ciência da Informação, e percebe-se como a área se comporta em relação à autoria e coautoria de artigos.

Periódicos

Em relação aos periódicos que mais publicaram sobre o tema, identificaram-se 70 títulos (ver gráfico 5). Sendo que 10 periódicos (14,3%) publicaram 48 artigos (37,8%), ou seja, em apenas 10 títulos (Tabela 4), temos quase 40% de toda a literatura publicada. Um único periódico publicou 11 artigos (8,7%), em segundo lugar, aparecem dois com seis artigos (4,72%) em cada, em terceiro lugar, quatro periódicos (3,15%) com quatro artigos publicados e, em quarto lugar, aparecem três periódicos, que publicaram três artigos cada. Se analisar a lista completa de título de periódicos, percebe-se que 14 periódicos publicaram juntos 64 artigos (46,72%) do total recuperado.

Periódico	ISSN	N	% N	Número de Artigos Publicados	Artigos Acumulados	% Artigos Acumulados
Journal of Librarianship and Scholarly Communication	2162-3309	1	1,4%	11	11	8,7%
Library Hi Tech	0737-8831	2	2,9%	6	17	13,4%
The Eletronic Library	0264-0473	3	4,3%	6	23	18,1%
Aslib Journal of Information Management	2050-3806	4	5,7%	4	27	21,3%
Ciência da Informação	1518-8353	5	7,1%	4	31	24,4%
Program: eletronic library and information systems	0033-0337	6	10%	4	39	30,7%
Informação & Informação	1981-8920	7	11,4%	3	42	33,1%
Online Information Review	1468-4527	8	12,9%	3	45	35,4%
Revista Digital de Biblioteconomia & Ciência da Informação	1678-765X	9	14,3%	3	48	37,8%

Tabela 4. Lista dos periódicos que mais publicaram artigos no tema.

Ao analisar a totalidade dos 127 artigos recuperados, dispersos nos 70 diferentes periódicos (Figura 5), 41 periódicos (59%) aparecem com um artigo, 19 periódicos (27%) com dois artigos, três periódicos (4%) com três artigos, quatro periódicos (6%) com quatro artigos, dois periódicos (3%) com seis artigos e um periódico (1%) com 11 artigos.

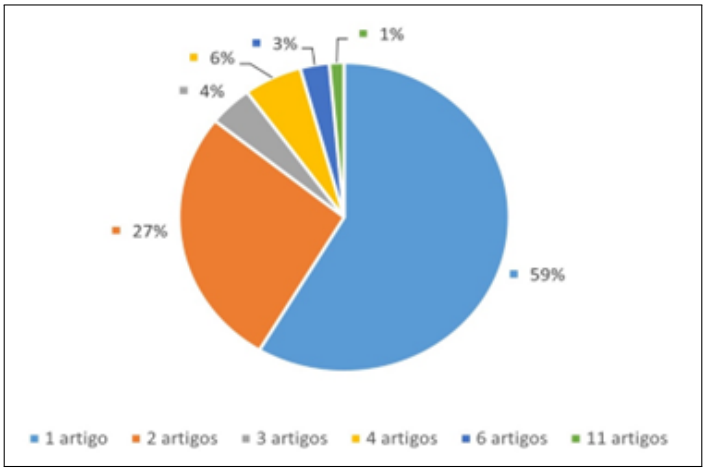

Figura 5. Distribuição da produção por periódicos.

Percebe-se que três primeiros periódicos são responsáveis por 23 artigos (Tabela 4), ou 18% de todos os artigos recuperados, sendo esses três periódicos internacionais.

Artigos

Em relação aos idiomas e anos dos artigos (Figura 6), percebe-se que no idioma inglês começou-se a publicar após 2005. Porém recuperou-se um artigo de 1998, que tem por assunto a gestão de dados digitais, deixando este na lista de artigos para conhecer os anseios e a perspectiva histórica do tratamento de dados digitais, mesmo sendo um artigo que pode contribuir pouco para a realidade atual. Os trabalhos mais antigos, de 2005 a 2008, referem-se a assuntos como padrões de metadados, preservação digital e protocolo OAI-PMH. Sendo o primeiro trabalho específico sobre os repositórios de dados de 2009.

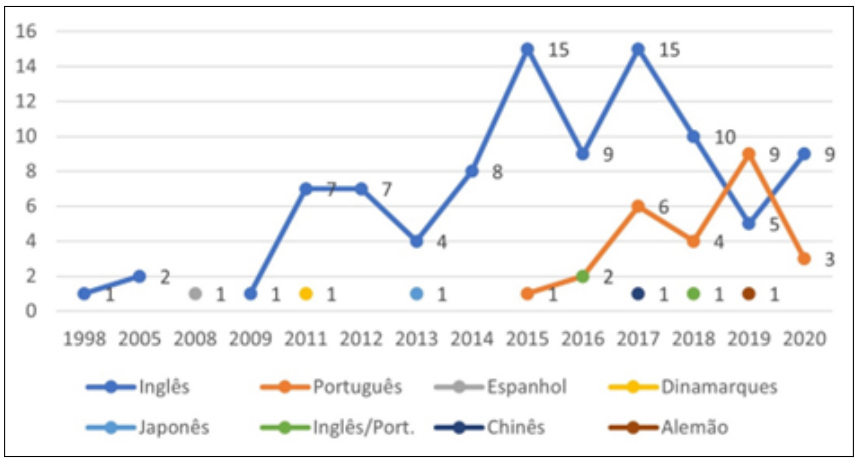

Figura 6. Distribuição da produção por periódicos.

Percebe-se (Figura 6) que, de 2011 a 2017, um grande crescimento da literatura em língua inglesa, tendo uma queda em 2018 e 2019, recuperando um pouco no ano de 2020. Porém, em relação ao português, o primeiro artigo recuperado é de 2015, tendo um crescimento em 2019, atingindo nove artigos, superando assim os recuperados em inglês no referido ano. Outros idiomas foram espanhol (2008), dinamarquês (2011), japonês (2013), chinês (2017) e alemão (2019).

Palavras-chave	Número de Artigos	%	% Acumulada
3 palavras	13	12,1%	12,1%
4 palavras	26	24,3%	36,4%
5 palavras	32	29,9%	66,4%
6 palavras	26	24,3%	90,7%
7 palavras	5	4,7%	95,3%
8 palavras	3	2,8%	98,1%
9 palavras	1	0,9%	99,1%
14 palavras	1	0,9%	100%
Total de artigos	**107**		

Tabela 5. Número de palavras-chaves.

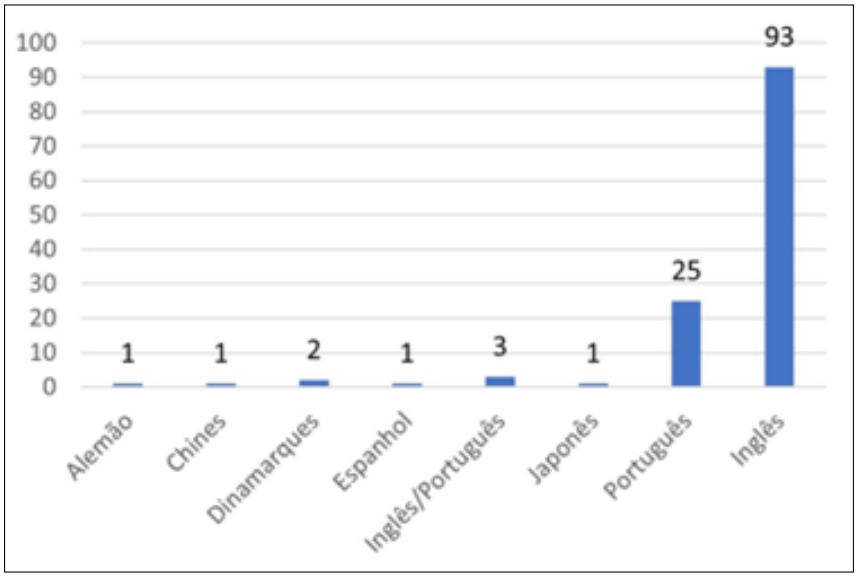

Figura 7. Produção por idiomas.

Outro dado é que, dos 127 artigos recuperados (Figura 7), 93 eram originalmente em inglês e 25 no português, porém uma revista brasileira disponibiliza o artigo integral em formato *Portable Document Format* (PDF) em português e inglês. Dessa forma, foi analisado em separado. Percebe-se que o periódico foi o único brasileiro da área que disponibiliza seus textos em dois idiomas.

Palavras-chave

Observou-se que alguns dos artigos em linha inglesa não traziam as palavras-chave presentes após o resumo ou em outra parte do texto. Assim, dos 127 artigos somente 107 tinham palavras-chave (Tabela 5), ou seja, em 84,25% dos artigos. Nestes, encontraram 542 termos, portanto tendo uma média de 5,06 palavras-chave por artigo, a média é de cinco palavras-chave.

Dentre as 542 palavras-chaves (tabela 6), identificaram-se 292 diferentes termos para representar os assuntos. Alguns foram agrupados e padronizados, como o uso de plural ou singular ou agrupando termos que representam o mesmo objetivo/assunto, por exemplo, "repositório de dados de pesquisa" ou "repositório de dados científicos", que foram agrupados em "repositório de dados". As palavras-chave que mais se repetiram foram:

	Palavras	Repetição
1	Repositório de dados	43
2	Dados de pesquisa	32
3	Gestão de dados de pesquisa	23
4	Compartilhamento de dados	15
5	Metadados	15
6	Curadoria de dados	10
7	Repositórios institucionais	9
8	Acesso aberto	8
9	Dados abertos	8
10	Reuso de dados	8
11	Ciência aberta	7
12	Curadoria digital	7
13	Acesso livre	6
14	*e-Science*	6
15	Dados	5
16	Gerenciamento de dados de pesquisa	5
17	Repositório	5
18	Bibliotecas acadêmicas	4
19	Bibliotecas digitais	4
20	Plano de gerenciamento de dados	4
21	Preservação digital	4
22	Serviços de dados	4
	Total de palavras	**232**

Tabela 6. Palavras-chave mais utilizadas.

As 23 palavras (Tabela 6) representam 53,87% de todas as recuperadas, ou seja, essas somadas aparecem 232 vezes. Pode-se concluir que, assim, são as palavras que devem ser usadas nas bases de dados para recuperação de artigos sobre o tema. Comparando com a Tabela 1 (confeccionada antes das buscas nas bases de dados), aparecem quase todos os termos, menos os ligados aos direitos de propriedade intelectual: direitos autorais e licenças.

CONSIDERAÇÕES FINAIS

Neste artigo, percebeu-se que os repositórios de dados são um tema em crescimento na literatura de Ciência da Informação. Percebe-se que começou a ser estudado no Brasil desde 2015, porém em países desenvolvidos, em 2009. Mesmo aparecendo oito autores brasileiros mais produtivos sobre o tema (dentre os 19 autores mais produtivos, ver Tabela 3), estes publicaram em coautoria, e observa-se que se recuperaram somente 25 artigos em português e 93 em inglês e três em inglês/português (sendo em periódico brasileiro), sendo que quatro das bases de dados pesquisadas são nacionais (SciELO, Portal de Periódicos da CAPES, BRAPCI, BDTD) e uma internacional (LISA). Assim, recuperaram-se mais artigos em línguas estrangeiras, confirmando Oliveira e Silva (2016, p. 6), que a realidade brasileira voltada para a ciência aberta e dados de pesquisa encontra-se em um estágio incipiente.

Convém reforçar que a gestão dos repositórios de dados de pesquisa deverá ficar a cargo dos "bibliotecários de dados" (Semeler, 2019). São bibliotecários que se especializarão em realizar a curadoria dos dados de pesquisa. Assim, a Biblioteconomia é chamada a desenvolver novas competências de seus profissionais para trabalharem com essas novas práticas do fazer científico, segundo Semeler (2019, p. 138), "[...] os repositórios de dados de pesquisa são parte da ciberinfraestrutura de *e-Science* e que podem ser vistos como uma das principais áreas de atuação para bibliotecários de dados preocupados em gerar serviços ou produtos de dados em bibliotecas." Portanto, verifica-se a necessidade de mais estudos no Brasil (eventos e pesquisa na pós-graduação) sobre implantação dos repositórios de dados para a capacitação de profissionais na gestão e implantação dessa tecnologia e seus serviços.

Essa constatação é um importante ponto de reflexão para a área de Ciência da Informação: o que desejamos quando propomos um serviço de repositório de dados? Quais são as características funcionais definidoras desse tipo de serviço? Que requisitos tecnológicos uma plataforma precisa ter para que seja possível a implantação de um repositório de dados? A existência de serviços de repositórios convencionais nomeados como "repositórios de dados" é um fato que ainda denota a necessidade do debate e a disseminação das técnicas de desenvolvimento desse tipo de serviço.

REFERÊNCIAS

Cervo, A. L., & Bervian, P. A. (2002). *Metodologia científica* (5a. ed.). Prentice Hall: São Paulo.

Dudziak, E. (2016). *Dados de pesquisa agora devem ser armazenados e citados.* São Paulo: USP. Recuperado de http://www.sibi.usp.br/?p=6189

Henning, P. C., Ribeiro, C. J. S., Sales, L. F., Moreira, J. L. R., & Santos, L. O. B. S. (2019). Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, *14*(3), 175–192. doi: 10.22478/ufpb.1981-0695.2019v14n3.46969

Monteiro, E. C. d. S. d. A. (2017). *Direitos autorais nos repositórios de dados científicos: análise sobre os planos de gerenciamento dos dados* (Dissertação de mestrado, Universidade Estadual Paulista Júlio de Mesquisa Filho, Marília, SP, Brasil). Recuperado de http://hdl.handle.net/11449/149748 (Tese de Doutorado)

Moreno, F. P. (2018). Repositórios de dados de pesquisa na espanha: breve análise. *Encontros Bibli: revista eletrônica de Biblioteconomia e Ciência da Informação*, *23*(53), 52–63. doi: 10.5007/1518-2924.2018v23n53p52

Oliveira, A. C. S., & Silva, E. M. (2016). Ciência aberta: dimensões para um novo fazer científico. *Informação & Informação*, *21*(2), 5–39. doi: 10.5433/1981-8920.2016v21n2p5

Pavão, C. G., Rocha, R. P. d., & Gabriel Junior, R. F. (2018). Proposta de criação de uma rede de dados abertos da pesquisa brasileira. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, *16*(2), 329–343. Recuperado de https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8651180/pdf

Pavão, C. G., Vanz, S. A. d. S., Caregnato, S. E., Moura, A. M. M. d., Passos, P. C. S. J., Gabriel Junior, R. F., ... Borges, E. N. (2019). *Acesso aberto a dados de pesquisa no brasil: políticas para repositórios de dados de pesquisa.* Recuperado de http://hdl.handle.net/20.500.11959/1263

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico* (2a. ed.). Novo Hamburgo: Feevale.

Rocha, D., & Schmidt, C. (2011). Jim gray e a escience: um método científico transformado. In T. Hey, S. Tansley, & K. Tolle (Eds.), *O quarto paradigma: descobertas científicas na era da e-sience* (p. 17–29). Microsoft. Recuperado de https://brapci.inf.br/index.php/res/download/125247

Sanchez, F. A., Vechiato, F. L., & Vidotti, S. A. B. G. (2019). Encontrabilidade da informação em repositórios de dados: uma análise do dataone. *Informação & Informação*, *24*(1), 51–79. doi: 10.5433/1981-8920.2019v24n1p51

Sayão, L. F., & Sales, L. F. (2015). *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.* Rio de Janeiro: CNEM/IEN. Recuperado de http://www.icb.usp.br/~sbibicb/images/guia%20gestaoPDF/Guia%20de%20gestao%20dados%20de%20pesquisa.pdf

Sayão, L. F., & Sales, L. F. (2016). Algumas considerações sobre os repositórios digitais de dados de pesquisa. *Informação & Informação*, *21*(2), 90–115. doi: 10.5433/1981-8920.2016v21n2p90

Semeler, A. R. (2019). *Ciência da informação em contextos de e-science: bibliotecários de dados em tempos de data science* (Tese de doutorado, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil). Recuperado de https://repositorio.ufsc.br/handle/123456789/185593 (Tese de Doutorado)

Silva, F. C. C. d. (2019). *Gestão de dados científicos.* Rio de Janeiro: Interciência.

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). *Primer on data management: what you always wanted to know.* California: CDL. Recuperado de http://escholarship.org/uc/item/7tf5q7n3#page-1

Universidade de São Paulo. (s.d.). *Gestão de dados científicos.* São Paulo: USP. Recuperado de http://prp.usp.br/gestao-de-dados-cientificos/?codmnu=9979

Como citar este artigo (APA):
Matos, I. Y. R., Prado, J. K. do, Ribeiro Júnior, D. I., & Argino, J. D. (2023). Produção científica sobre repositório de dados de pesquisa: representações autorais e temáticas. *AtoZ: novas práticas em informação e conhecimento, 12*, 1 – 13. Recuperado de: http://dx.doi.org/10.5380/atoz.v12.85255

NOTAS DA OBRA E CONFORMIDADE COM A CIÊNCIA ABERTA

CONTRIBUIÇÃO DE AUTORIA

Papéis e contribuições	Igor yuri Ramos Matos	Divino Ignacio Ribeiro Júnior	Jorge Kroll do Prado	Julibio David Argino
Concepção do manuscrito	X			
Escrita do manuscrito	X	X		
Metodologia	X		X	
Curadoria dos dados	X			X
Discussão dos resultados	X	X		X
Análise dos dados	X	X	X	

Disponibilidade de Dados Científicos da Pesquisa

Os conteúdos subjacentes ao texto da pesquisa estarão disponíveis no momento da publicação do artigo.

EQUIPE EDITORIAL