



Encontros Bibli: revista eletrônica de biblioteconomia e
ciência da informação

ISSN: 1518-2924

adilson.pinto@ufsc.br

Universidade Federal de Santa Catarina
Brasil

da SILVEIRA, Lúcia; Dall'Agnol BARBOSA, Amanda;
Klanovicz FERREIRA, Manuela; CAREGNATO, Sônia Elisa

CITAÇÃO DE DADOS CIENTÍFICOS: SCOPING REVIEW

Encontros Bibli: revista eletrônica de biblioteconomia
e ciência da informação, vol. 25, 2020, -, pp. 1-31

Universidade Federal de Santa Catarina
Brasil

DOI: <https://doi.org/10.5007/1518-2924.2020.e72153>

Disponível em: <https://www.redalyc.org/articulo.oa?id=14763386032>

- ▶ Como citar este artigo
- ▶ Número completo
- ▶ Mais informações do artigo
- ▶ Site da revista em redalyc.org

UAEM redalyc.org


Sistema de Informação Científica Redalyc


Rede de Revistas Científicas da América Latina e do Caribe, Espanha e Portugal


Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa
acesso aberto


CITAÇÃO DE DADOS CIENTÍFICOS: SCOPING REVIEW


Scientific Data Citation: Scoping Review

Lúcia da SILVEIRA
Doutoranda em Comunicação
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil
luciadasilveiras@gmail.com
<http://orcid.org/0000-0003-1118-2121> 

Manuela Klanovicz FERREIRA
Mestra em Computação
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil.
manuelakf@cpd.ufrgs.br
<http://orcid.org/0000-0002-9089-7725> 

Amanda Dall'Agnol BARBOSA
Graduanda em Biblioteconomia
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil
amandadallagnolbarbosa@gmail.com
<http://orcid.org/0000-0003-1052-5196> 

Sônia Elisa CAREGNATO
Professora da Faculdade de Biblioteconomia e Comunicação e
do Programa de Pós-Graduação em Comunicação
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil
sonia.caregnato@ufrgs.br
<http://orcid.org/0000-0002-5676-2763> 

A lista completa com informações dos autores está no final do artigo 

RESUMO

Objetivo: Para acompanhar a evolução dos estudos referentes a dados científicos, investigou-se o significado das citações a eles, buscando responder: 1) Quais as motivações dos pesquisadores para citar dados científicos? 2) Quais as práticas de citação de dados apresentadas nas áreas cobertas pelo presente estudo? 3) Quais as análises métricas para citação de dados?

Método: Caracteriza-se como uma pesquisa do tipo qualitativa e descritiva, uma revisão de literatura do tipo *Scoping Review*, com busca às bases de dados *Emerald*, *LISA*, *LISTA*, *Scopus* e *Web of Science*.

Resultados: Como motivação, identificaram-se estudos sobre a correlação entre o incremento de citações e as publicações tradicionais. Muitas destas, ao citarem os dados que as embasavam, confirmaram a correlação, outras não, surgindo também a hipótese de causa comum: qualidade da pesquisa associada a mais recursos. Quanto às práticas, a comunidade está ciente de que as citações atuais de dados não estão padronizadas, o que evidencia uma tendência para a adoção de um padrão de citação que atenda às demandas de diferentes tipos de dados. Essa falta de padrão dificulta a análise métrica de citação de dados científicos, que ainda precisa ser explorada em pesquisas, tendo em vista que há uma repetição em utilizar as mesmas técnicas da citação tradicional para essa nova fonte de informação.

Conclusões: Promover o avanço da ciência é a principal vantagem de disponibilizar dados, mas existem dificuldades técnicas e de atribuição de crédito que precisam ser enfrentadas em conjunto por pesquisadores, instituições, agências de fomento, repositórios de dados e equipes editoriais de publicações.

PALAVRAS-CHAVE: Citação de dados científicos. Métricas. *Scoping review*.

ABSTRACT

Objective: This paper investigates the meaning assigned to data citation in order to follow the evolution of studies related to data citation, it tries to answer: 1) What are the motivations of researchers to cite scientific data?; 2) What are the data citation practices presented by the areas covered by this study?; 3) What are the metric analysis for data citation?

Methods: It is a qualitative and descriptive research, being a scoping review of literature, by searching the Emerald, LISA, LIST, Scopus and Web of Science databases.

Results: The studies investigated the correlation of citations increment to traditional publications by citing the data that supported them, many studies confirmed the correlation, others did not, and a common cause hypothesis arose: research quality associated with more resources. As for practices, the community is aware that current citations to data are not standardized, and there is a tendency to adopt a citation standard that meets the demands of different types of data. This lack of standard hinders the metric analysis of citation to scientific data that still needs to be explored in research, given that there is a repetition in using the same techniques of traditional citation for this new source of information.

Conclusions: Promoting the progress of science is the main advantage in making data available, but there are credit and technical difficulties that need to be tackled together by researchers, institutions, funding agencies, data repositories, and publishing editorial teams.

KEYWORDS: Scientific data citation. Metrics. Scientific data management. Scoping review.

1 INTRODUÇÃO

O compartilhamento e reuso de conjuntos de dados científicos está em plena ascensão, essencialmente pelo fato de sua citação ser um requisito nas publicações a que estão relacionados. Por meio da citação é preciso que os conjuntos de dados digitais referenciados sejam facilmente descobertos e inequivocamente localizáveis, por humanos e máquinas, na vastidão da internet, além de dar a valorização correta a cada pessoa e instituição que contribuiu para a coleta e curadoria dos dados.

Esse tema tornou-se agenda dos cientistas de distintas áreas, pois por intermédio dele busca-se como objetivo principal tornar acessíveis e públicas as pesquisas científicas em todo o seu ciclo de vida, de modo a promover a confiabilidade das informações por meio do aumento de sua reprodutibilidade e reuso dos dados. Mais do que isso, o reuso redesenha o potencial cognitivo dos dados, no qual o valor atribuído à pesquisa é vinculado ao potencial de seus dados serem reinterpretados em suas áreas de origem ou em outras áreas, assim como em diferentes contextos, estabelecendo novos padrões de socialização e de trabalho cooperativo, independentemente de barreiras geográficas ou disciplinares (SAYÃO; SALES, 2014).

A citação na ciência é uma representação simbólica de um conhecimento construído coletivamente e, portanto, um bem comum a qualquer cidadão. Citar é respeitar, valorizar, homenagear, criticar, refutar (PIWOWAR; DAY; FRIDSMA, 2007) as construções do conhecimento por meio de um fragmento de texto, imagem, dados numéricos, protocolos, códigos (SILVA, 2019), ou seja, de certa expressão que tenha algum sentido para seres humanos, isto é, um registro de informação científica. Os dados científicos, também denominados “dados de pesquisa”, são um subconjunto dessa informação, definidos pela *Organisation for Economic Co-operation and Development* (OCDE, 2007, tradução nossa) como “registros factuais usados como fonte primária para a pesquisa científica e que são comumente aceitos pelos pesquisadores como necessários para validar os resultados do trabalho científico”.

A citação dos conjuntos de dados ocorre quando o autor decide utilizá-los ou, possivelmente, reutilizá-los, estes sendo ou não do próprio autor ou estando ou não públicos. Nesse sentido, os dados científicos precisam ser autênticos e consistentes, bem como devem seguir padrões mínimos como os princípios FAIR¹ (COUSIJN et al., 2018),

¹ FAIR significa: a) localizáveis (*Findable*): com metadados ricos e indexáveis, bem como um identificador persistente; b) acessíveis (*Accessible*): recuperáveis por meio de um protocolo padronizado, aberto e

por meio dos quais são definidos requisitos mínimos que também norteiam a evolução dos padrões e das ferramentas de citação de dados. A atribuição de um identificador persistente a um conjunto de dados facilita a atribuição do crédito aos pesquisadores e às instituições que produzem e mantêm os dados primários, além de garantir a rastreabilidade do objeto (NOVACESCU et al., 2018) e a identificação unívoca do conjunto de dados (HERTERICH; DALLMEIER-TIESSEN, 2016).

A exigência de publicar os dados científicos é realidade em diferentes contextos, prevalecendo as iniciativas internacionais. Por exemplo, agências de fomento como a *National Science Foundation* (NSF), a *National Institutes of Health* (NIH), o *National Cancer Institute* (NCI) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) exigem a submissão de um Plano de Gestão de Dados (PGD) em conjunto com a proposta de pesquisa a ser avaliada para fomento. Sendo assim, plataformas foram criadas com serviços de gestão e publicação dos dados de pesquisa, como FigShare (2019), Zenodo (2019) e Datadryad (2019), estas conveniadas ao DataCite (2019), possibilitando que os conjuntos de dados associados a esses repositórios recebam um DOI.

Outro avanço são as ferramentas de descoberta de dados, tais como a *Data Citation Index* (DCI), da *Web of Science* (WoS), criada em 2012 (PAVLECH, 2016); a *Mendeley Data*, originada da parceria da *Elsevier* com o *DataCite* e com numerosos repositórios de dados a fim de ligar os artigos do *ScienceDirect* aos respectivos dados (BELTER, 2014); e o *Google Dataset Search* (GDS), mais recente, inaugurado em outubro de 2018, o qual relaciona os conjuntos de dados com os artigos do *Google Scholar* (GS). Entre estas, o DCI e o GDS também fornecem a quantidade de citações dos dados feita em publicações tradicionais presentes no *Web of Science* e no *Google Scholar*, respectivamente.

Diante dessa perspectiva, atesta-se uma preocupação por parte de distintas organizações com os vínculos entre as produções mais tradicionais, tais como artigos, livros, dissertações e teses, e seus conjuntos de dados científicos, e, conseqüentemente, em como utilizar esses vínculos de modo a garantir a transparência e a credibilidade das pesquisas. Dessa maneira, para acompanhar a evolução do cenário dos estudos

gratuito, com possibilidade de autenticação e níveis de acesso; c) interoperáveis (*Interoperable*): os metadados devem utilizar uma linguagem formal, acessível e compartilhada, além de vocabulários FAIR; d) reutilizáveis (*Re-usable*): com metadados ricamente descritos com atributos precisos e relevantes, tais como a precedência detalhada dos dados e a licença de uso clara e acessível, especificados em padrões estabelecidos pela comunidade (WILKINSON, et al., 2016).

relacionados a esse assunto, pretende-se neste artigo investigar qual o significado das citações a dados científicos, com a finalidade de buscar respostas para as seguintes perguntas: 1) Quais as motivações dos pesquisadores para citar dados científicos? 2) Quais as práticas de citação de dados apresentadas pelas áreas cobertas pelo presente estudo? 3) Quais as análises métricas para citação de dados?

2 OPÇÕES METODOLÓGICAS

Este artigo caracteriza-se como uma pesquisa do tipo qualitativa e descritiva, a qual emprega procedimentos de um *scoping review*, com base no manual disponível no *The Joanna Brigs Institute for Scoping Reviews* (JBI, 2020). *Scoping Review* é um tipo de levantamento bibliográfico cujo propósito é mapear estudos primários, artigos de pesquisa ou revisões em busca de evidências para responder a uma pergunta ou perguntas subjacentes relacionadas a um fenômeno. A técnica utiliza protocolo para estruturar a pergunta de investigação, a busca em fontes de informação, a seleção de artigos e a aplicação de filtros nos resultados, facilitando a reprodução da pesquisa.

A pergunta pode ser delineada por meio do mnemônico PCC: População, Conceito e Contexto. Seguindo o PCC, considera-se como (P) os pesquisadores, evidenciando as suas motivações; (C), conceito, como os dados científicos – entretanto, essa exploração conceitual não foi realizada –; por fim, (C), contexto, em que nesse caso foi situado como se dá a citação, as referências e as metodologias usadas para análise de citação de dados de acordo com os estudos encontrados. Com base no PCC, foram realizados os testes de estratégias de busca no período de maio a junho de 2019 com termos variantes, a fim de reconhecer os critérios de busca de cada base, bem como a consistência e a relevância dos resultados. Inicialmente, foram selecionados os termos que seriam utilizados na estratégia de busca (Apêndice A). Posteriormente, foi compartilhado o resultado com uma especialista na área com a finalidade de aferir a aderência dos artigos à estratégia e ao propósito da pesquisa. As bases *Scientific Electronic Library Online* (SciELO), Base de Dados em Ciência da Informação (BRAPCI) e Portal brasileiro de publicações científicas em acesso aberto - OASIS.br foram excluídas da seleção para que fosse mantido o mesmo padrão de estratégia em todas as fontes de informação, visto que estas não permitiam o uso da estratégia de busca composta, retornando resultados infiéis aos termos utilizados, ou seja, sem relevância. Nessa fase, constatou-se que a inclusão

do termo “*reuse*” foi pouco efetiva, pelo fato de retornar resultados irrelevantes para o objetivo da pesquisa, pois os artigos recuperados lidavam com questões relativas à gestão e ao compartilhamento de dados, e, portanto, estavam pouco associados ao termo “citação”. Por isso, o termo “*reuse*” foi excluído da estratégia final. Após esses testes, foi realizada a coleta final em 24 de junho de 2019, sendo então aberta a cobertura temporal para os testes e para a coleta final, ou seja, os artigos foram inseridos até o dia da realização da busca, para se obter a maior amplitude sobre o tema. Isso se justifica pelo conteúdo tratado ser relativamente novo na literatura científica, portanto, optou-se por essa abrangência aberta. No apêndice A, estão listadas as estratégias usadas nas buscas e as quantidades de artigos recuperados nas respectivas bases de dados, a saber: *Emerald*, *LISA*, *LISTA*, *Scopus* e *Web of Science*.

Uma limitação deste trabalho é o fato de tratar apenas de assuntos relacionados à citação dos conjuntos de dados, não sendo investigado o tratamento dos dados científicos para a formação desses conjuntos.

Ao final da busca, foi obtido um total de 84 referências, das quais foram removidas 29 duplicatas – 27 com uma ferramenta automática do *Mendeley* e 2 removidas manualmente pelo mesmo motivo –, restando 55 artigos. Para compor o *corpus* de análise, os artigos foram filtrados em duas etapas:

a) **Primeira etapa** – a seleção dos manuscritos ocorreu a partir da leitura de metadados (títulos, resumos, autores, palavras-chave) de todos os documentos, excluindo-se os que não tinham de fato relação com o tema de pesquisa, de modo a atender aos critérios de exclusão do Quadro 1. Para a classificação de relevância, foram adotadas as expressões “muito relevante”, “relevante”, “em dúvida” e “excluído”. Para ser considerado relevante, o artigo precisaria responder a uma das perguntas deste estudo; respondendo a mais de uma, seria considerado muito relevante. Foram excluídos 11 artigos, por não serem relevantes.

Quadro 1 -- Critérios de inclusão e exclusão de busca e seleção

Critérios de inclusão	Critérios de exclusão
1) Inglês, espanhol, português 2) Acesso completo 3) Artigos de conferências 4) Revisado por pares 5) Termos presente no título 6) Resumo representativo 7) Leitura técnica para ver se respondia alguma pergunta	1) Duplicidade de artigo 2) Demais idiomas (chinês, francês, alemão...) 3) Não revisado por pares 4) Título não representativo 5) Resumo não representativo

Fonte: das autoras (2019).

b) Segunda etapa – leitura completa dos artigos e elaboração de fichamentos descritivos de acordo com as perguntas da presente pesquisa, utilizando-se uma planilha compartilhada para permitir a organização simultânea pelas autoras das informações dos documentos previamente selecionados, com vistas a eliminar artigos irrelevantes para esse estudo. Nessa etapa, do conjunto de 44 artigos restantes da Etapa 1, foram excluídos 14 artigos por irrelevância para esse estudo.

Como resultado, foram obtidos 30 artigos relevantes, distribuídos entre 1995 e 2019, cujos conteúdos foram categorizados, descritos e sintetizados. Esse tratamento foi aplicado em cada artigo, sendo a escrita dos resultados organizada por ordem cronológica, respeitando-se em primeiro plano as evidências relacionadas aos três questionamentos da presente pesquisa, sumarizadas no Quadro 2.

Quadro 2 – Lista de 30 artigos selecionados

#	Ano	Autores	Motivações para citar conjunto de dados	Práticas de citação de dados	Métodos de análise de citação de dados
1	1995	Sieber; Trumbo	✓	✓	✓
2	2007	Piwowar; Day; Fridsma	✓		✓
3	2012	Callaghan; Lowry; Walton	✓	✓	
4	2012	Mayernik	✓		
5	2012	Mooney; Newton		✓	
6	2013	Altman; Corsas		✓	
7	2013	Piwowar; Vision	✓		✓
8	2013	Simons; Visser; Searle		✓	
9	2014	Belter	✓	✓	✓
10	2014	Force; Robinson		✓	
11	2013	Mooney		✓	
12	2015	Altman et al.	✓	✓	
13	2015	Henderson; Kotz	✓	✓	✓
14	2015	Mathiak; Boland		✓	✓
15	2015	Pröll; Rauber		✓	
16	2016	Buneman; Davidson; Frew		✓	
17	2016	Herterich; Dallmeier-Tiesse			✓
18	2016	Onyancha			✓
19	2016	Peters et al.			✓
20	2016	Robinson-García; Jiménez-Contreras; Torres-Salinas	✓	✓	✓
21	2016	Zwölf; Moreau; Dubernet		✓	
22	2017	Park; Wolfram	✓		✓
23	2018	Li; Chen	✓		

#	Ano	Autores	Motivações para citar conjunto de	Práticas de citação de dados	Métodos de análise de citação
24	2018	Novacescu et al.	✓		
25	2018	Park; You; Wolfram	✓		✓
26	2018	Silvello, Gianmaria	✓	✓	
27	2018	Zhao; Yan; Li			✓
28	2018	Cousijn et al.	✓	✓	
29	2019	Park; Wolfram		✓	
30	2019	Zwölf et al.	✓	✓	

Fonte: dados desta pesquisa (2019).

O apêndice B lista os 25 artigos que foram excluídos por não serem considerados relevantes para a presente pesquisa. O retrato da literatura (Quadro 2) é abordado na apresentação dos resultados a seguir.

3 APRESENTAÇÃO DOS RESULTADOS

O mapa de produção científica a respeito das citações a conjuntos de dados mostrou que a literatura é recente, tendo em vista que 28 títulos foram encontrados nos últimos sete anos (2012-2019) e, antes desse período (1995-2007), apenas dois títulos. Dos 30 artigos que compuseram o *corpus* desta pesquisa, 18 foram publicados na área de Ciência da Informação, o que evidencia o domínio desse campo sobre o tema de estudo em questão.

Dos artigos selecionados, 16 apresentam conteúdos relevantes sobre a motivação para citar conjuntos de dados, 19 trabalhos são relacionados às práticas de citação e 13 referem-se aos métodos de análise de citação de dados.

Em relação aos motivos para citar o conjunto de dados científicos, destaca-se que os artigos recuperados não são oriundos de pesquisa aplicada ou experimental, contendo apenas argumentos dos autores que mostravam vantagens, desvantagens, contextos e experiências a respeito desse assunto. Ainda que tenha essa limitação, a descrição das principais fundamentações abordadas foi mantida.

As pesquisas associadas às práticas de citação de conjunto de dados possuem foco nas abordagens metodológicas aplicadas, e analisam, por exemplo, os elementos utilizados na elaboração das referências para conjuntos de dados científicos.

Quanto aos tipos de análises métricas para citação de dados, a literatura caracterizou-se, em sua maioria, como pesquisa aplicada, investigando se as

metodologias tradicionais de análises de citação poderiam ser usadas nesse tipo de produção.

As próximas seções apresentam as evidências relacionadas às três perguntas para compor a presente revisão de literatura.

3.1 Motivações para citar dados científicos

A motivação para citar um conjunto de dados científicos pode estar atrelada a diversos fatores que ultrapassam a tomada de decisão do ato de citar dados, de modo que vão desde a disposição e o comprometimento em compartilhar até a confiança de reusar os dados de terceiros. Nesse sentido, a presente seção intenta destacar motivações, bem como apontar vantagens e desvantagens encontradas na literatura a respeito dessa discussão.

A motivação principal ao citar dados científicos é a possibilidade de obter mais colaboração, rastreabilidade e, conseqüentemente, impacto e visibilidade (PIWOWAR; DAY; FRIDSMA, 2007; CALLAGHAN; LOWRY; WALTON, 2012; PIWOWAR; VISION, 2013; NOVACESCU et al., 2018; COUSIJN et al., 2018). Isso está amparado na justificativa de que a reutilização dos dados aumenta as citações, como relatam algumas pesquisas das áreas da Saúde, da Ciência da Informação, da Física, da Oceanografia, das Engenharias, da Ciência da Computação e da área Interdisciplinar (PIWOWAR; DAY; FRIDSMA, 2007; PIWOWAR; VISION, 2013; BELTER, 2014; HENDERSON; KOTZ, 2015; ROBINSON-CARCÍA; JIMÉNEZ-CONTRERAS; TORRES-SALINAS, 2016; SILVELLO, 2018; PARK; YOU; WOLFRAM, 2018).

Para Piwowar e Vision (2013), publicar os dados antes do resultado em plataformas distintas das da publicação da pesquisa pode aumentar a recuperação da informação, dando maior visibilidade ao trabalho e, conseqüentemente, maior possibilidade de citação.

Piwowar e Vision (2013) revelam que as autocitações aos conjuntos de dados são mais frequentes nos dois primeiros anos e que há um aumento de citação de terceiros em um período de 3 a 6 anos a partir do momento em que o dado é publicado. Na Oceanografia, Belter (2014) verificou que mesmo os conjuntos de dados antigos continuam recebendo citações muito além do que costuma ser visto em artigos científicos tradicionais.

Na área da Saúde, Piwowar, Day e Fridsma (2007) pontuam que o fato de uma pesquisa receber mais citações pode não ser consequência de seus dados serem compartilhados; ao invés disso, podem ser resultados das mesmas causas. Exemplificam que, quando um estudo exige um ensaio clínico de grande volume de dados de alta qualidade, naturalmente a obra recebe mais citações devido a sua relevância médica e, isso posto, os pesquisadores estão mais inclinados a compartilhar seus dados do que estariam no caso de um ensaio pequeno, talvez em virtude de seus recursos mais abundantes ou para gerar maior confiança em relação aos resultados da pesquisa.

Piwowar e Vision (2013) apontam como uma vantagem da citação de dados a possibilidade de realizar novos estudos que não foram previstos na pesquisa que originou esses dados, dando-se preferência a esses estudos como fundamentações teóricas. É o que comprova Henderson e Kotz (2015), ao verificarem que dados disponibilizados foram reutilizados por diversas outras áreas além das de Oceanografia e Ciências Sociais.

No caso da Astronomia, uma motivação é a utilização de dados oriundos de telescópios de pesquisa que, devido à própria magnitude e complexidade dos dados, exige a colaboração entre pesquisadores e o respectivo reconhecimento por meio da citação dos dados (NOVACESCU et al., 2018).

Para Altman et al. (2015), a ligação dos artigos aos dados possibilita novas formas de publicação científica, promove a pesquisa interdisciplinar, fortalece a ligação entre política e ciência, como também diminui o custo da replicação e da continuidade de pesquisas prévias.

Autores da área de Biomedicina e Oceanografia estão mais propensos a compartilhar dados se houver a identificação de créditos explícitos, requerendo a citação formal, o que não ocorre com os casos de citações informais em seções de agradecimento (PARK; YOU; WOLFRAM, 2018). Isso pode acontecer porque essas áreas são mais estruturadas em repositórios de dados científicos e na infraestrutura de colaboração.

Diante das percepções desses autores, entende-se que o ato de citar dados científicos ultrapassa questões de ordem técnica, mobilizando os pesquisadores a colaborar, inovar e respeitar os direitos autorais de seus achados (Figura 1).

Figura 1 – Fatores motivacionais para citar dados científicos



Fonte: dados da pesquisa. Elaboração das autoras (2019). Ferramenta: Canva.

As variáveis inerentes à transparência dos dados (entre elas, a propriedade intelectual e a confidencialidade dos dados) podem ser também inibidoras do seu compartilhamento (PIWOWAR; DAY; FRIDSMA, 2007). Para que o pesquisador não publique seus dados científicos, devem ser considerados motivos como falta de recursos apropriados, volume de dados, tempo para formatação, tratamento, limpeza de confidencialidade dos resultados, direitos autorais das fontes dos dados, desidentificação dos dados e descrição de metadados, dificuldade em decidir quais dados serão liberados e, ainda, definir onde os dados serão publicados – evitando a transitoriedade do local de armazenamento, o uso indevido dos dados, bem como a falta de um sistema de recompensa mais estruturado (PIWOWAR; DAY; FRIDSMA, 2007; PARK; YOU; WOLFRAM, 2018).

Além disso, a transparência desses dados poderá causar receio do pesquisador quanto à própria credibilidade deles (PIWOWAR; DAY; FRIDSMA, 2007), pois uma pesquisa, quando contestada em reanálise, poderá surtir efeitos significativos à vida do pesquisador, desde a sua exclusão da comunidade científica até a perda de seu trabalho ou cargo. Esse risco requer dos pesquisadores um comprometimento ainda mais rigoroso com a conduta ética e íntegra de cientista.

Outros fatores que contribuem para que os pesquisadores mantenham seus dados invisíveis são relativos ao tempo, tanto o despendido pelo autor na disponibilização do dado, fazendo novas revisões ou rebatendo refutações (PIWOWAR; DAY; FRIDSMA, 2007), quanto o poupado pelo concorrente ao reutilizar os dados de terceiros (PARK;

WOLFRAM, 2017). Disponibilizar os dados de uma pesquisa robusta poderá ser o fracasso da vantagem competitiva, ou mesmo da possibilidade de futuras publicações com o mesmo conjunto de dados, ou ainda, patenteamento de um produto com base nesses dados (PIWOWAR; DAY; FRIDSMA, 2007). Ou seja, isso poderá dificultar atividades produtivas ou lucrativas.

Diante da perspectiva de o uso das citações de dados ser uma moeda de troca para recompensar autores, Belter (2014) reforça que a utilização exclusiva de indicadores bibliométricos de publicações tradicionais para avaliação da pesquisa pode levar à desvalorização de outros tipos de atividades que não resultem em publicações científicas formais, levando ao baixo interesse de pesquisadores em dedicar seus esforços a tarefas como a curadoria de dados científicos. O autor sugere a utilização de métricas alternativas como quantidade de *downloads*, discussão em redes sociais e social *bookmarking*.

As diferentes falas desses autores a respeito das fragilidades envolvidas na abertura dos dados para reutilização, a qual torna possível, consequentemente, a citação do dado, deixam muitas questões a serem aprofundadas em outros estudos. Suas pesquisas realçaram a necessidade de ter mais recursos, resistências quanto à credibilidade dos dados, entre outras questões apontadas na Figura 2.

Estratégias de recompensa para o engajamento dos pesquisadores na publicação de dados foram apontadas no estudo de Henderson e Kotz (2015). Para divulgar o propósito do repositório de dados *Community Resource for Archiving Wireless Data at Dartmouth* (CRAWDAD), foram realizados *workshops* em grandes congressos da área da Computação, os quais atualmente exigem o compartilhamento dos dados com os autores que submetem trabalhos para premiação. A fim de sensibilizar os autores quanto ao compartilhamento de seus dados, eles ganhavam um lagostim de brinquedo (identidade visual do repositório). A equipe do CRAWDAD auxiliava os autores que precisassem de orientações para preparar os dados finais para publicação. Nesse sentido, a proposta de ZWÖLF et al. (2019, p. 2) pretende implementar no VAMDC *Consortium* formas de colaboração dentro dos próprios *clusters* e recompensar os pesquisadores por suas contribuições.

Figura 2 – Fragilidades que desfavorecem a abertura e, consequentemente, a citação dos dados



Fonte: dados da pesquisa. Elaboração das autoras (2019). Ferramenta: Canva.

A citação de dados não afeta apenas os pesquisadores em si, mas também toda a rede estruturada da comunicação científica, como os periódicos, as agências de fomento e as instituições. Silvello (2018) afirma que são as agências de fomento que se preocupam em garantir a reprodutibilidade, não os autores.

Em um primeiro momento, é importante que a editoração de periódicos adote diretrizes para autores (SIEBER; TRUMBO, 1995) que incluam orientações claras a respeito do uso, reuso e compartilhamento dos dados, o que implica averiguar se o local em que os dados são armazenados segue padrões de qualidade, como os princípios FAIR, e se possui políticas de preservação. O periódico precisa ter como regra que: “um conjunto de dados não pode ser citado se não for devidamente arquivado” (MAYERNIK, 2012, p. 25) e preservado.

Uma inovação que surgiu em razão do contexto de compartilhamento de dados é a publicação em revistas de dados. Originadas pela necessidade de dar crédito aos autores dos dados (PARK; WOLFRAM, 2017) e destinadas a serem diários desse tipo de conteúdo (PARK; YOU; WOLFRAM, 2018), registram a evolução dos dados, já que se modificam rapidamente, relatam Zwölf *et. al* (2019) sobre a área de Física. Li e Chen (2018) constataram que o artigo de dados é um novo tipo de publicação que evidencia uma estrutura narrativa na organização de seu conteúdo. Os autores analisaram 40 dessas publicações e concluíram que estão internamente estruturadas em: histórico e resumo, métodos, registro de dados, validação técnica e notas de uso. Assim, esse tipo

de publicação poderá ser uma motivação a mais para os pesquisadores no momento de publicar seus dados, já que auxiliará tanto na ampliação das citações quanto no reconhecimento do trabalho realizado.

3.2 Práticas na citação de dados científicos: referência bibliográfica

Esta seção trata das experiências e investigações acerca de como a literatura científica tem se manifestado sobre as práticas de citação de dados, apontando as principais mudanças e evidências a esse respeito.

A citação tradicional não possui as mesmas características que a citação de conjuntos de dados (SILVELLO, 2018). Ambas têm um propósito comum, no entanto, se diferem epistemologicamente. É fundamental que os conjuntos de dados sejam citados diretamente para darem suporte aos resultados da pesquisa, somando-se às referências da literatura, mas sem substituí-las (COUSIJN et al., 2018).

Os princípios de citação para conjunto de dados foram descritos (CODATA, 2013) e posteriormente sumarizados pelo *Joint Declaration of Data Citation Principles* (JDDCP), sendo eles: 1) Importância: devem ser ofertados à citação de dados o mesmo valor e legitimação conferidos a outros registros de pesquisa; 2) Crédito e atribuição: todos os responsáveis pelos dados deverão ser reconhecidos na citação formal, considerando que um único estilo ou mecanismo de citação poderá ser insuficiente para a representação de diferentes conjuntos de dados; 3) Evidência: quando a pesquisa se baseia em dados, estes devem ser citados; 4) Identificação: exclusiva, persistente e acionável por máquina para cada conjunto de dados, utilizando padrões aceitos pela comunidade científica; 5) Acesso: as citações devem facilitar o acesso por humanos e máquinas aos metadados, aos conjuntos de dados e aos documentos associados a eles; 6) Persistência: URLs únicas para cada conjunto de dados independentemente da volatilidade dos dados; 7) Especificidade e verificação: citações e metadados devem garantir a identificação, o acesso, a procedência e a granularidade dos dados; 8) Interoperabilidade e flexibilidade: os métodos de citação devem garantir a especificação do conjunto de dados, seguindo um padrão de citação flexível a fim de representar melhor o conjunto de dados (FORCE11, 2014; ALTMAN, et al., 2015; HENDERSON; KOTZ, 2015; SILVELLO, 2018; COUSIJN et al., 2018).

Considerando esses princípios, deve-se ponderar a flexibilidade da descrição da obra para evitar atrapalhar a interoperabilidade dos sistemas. A padronização da

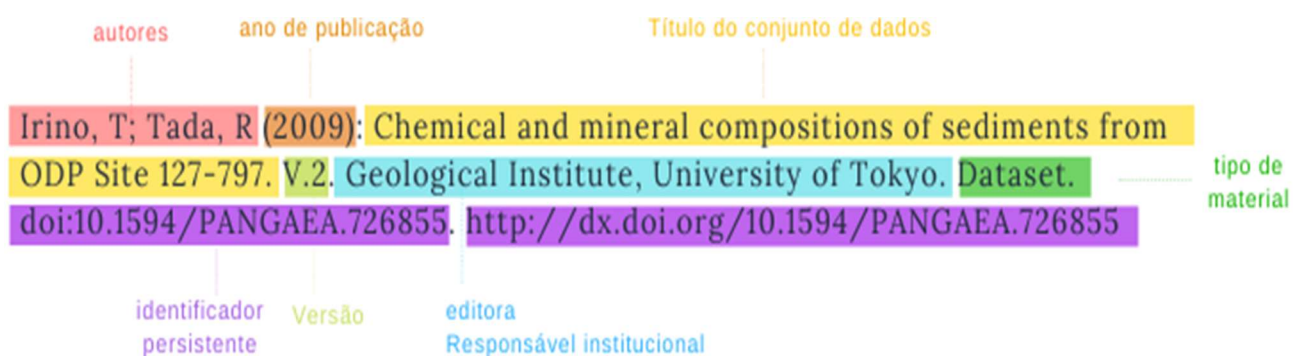
referência precisa ser suficiente para suprir a automatização da citação nas diferentes áreas do conhecimento, levando em conta a autoria, o significado e a especificidade.

Muitos dos trabalhos anteriores ao lançamento do *Joint Declaration of Data Citation Principles* (JDDCP), do FORCE11 (2014), analisaram os padrões, ainda informais, de como as esparsas citações de dados aparecem em publicações tradicionais como artigos de periódicos, analisando as características comuns de citações de dados em diferentes áreas do conhecimento (MOONEY; NEWTON, 2012; MOONEY, 2013; ALTMAN; CROSAS, 2013; SIMONS; VISSER; SEARLE, 2013; FORCE; ROBINSON, 2014).

Sieber e Trumbo (1995), por exemplo, definem os padrões de citação de dados considerando os elementos fundamentais de APA (1983) e Dodd (1979) para o conjunto de dados: nome do autor principal, título, data do levantamento ou lançamento do conjunto, produtor ou distribuidor dos dados, endereço destes, financiador do projeto, se o conjunto de dados pode ser lido por máquinas e se acompanha um livro de códigos. Mooney (2013) propôs os seguintes elementos: autor, data de publicação, título, incluindo aqui edição ou versão, editor ou distribuidor, localização eletrônica ou identificador.

Em 2016, a iniciativa do padrão *DataCite* (Figura 3) em incorporar novos descritores para conjunto de dados o tornou precursor no estabelecimento de um padrão de referência bibliográfica para dados científicos (SILVELLO, 2018). Anteriormente, a literatura apenas apontava autor, data de publicação, título, edição, versão, identificador persistente, tipo de recurso, editor, identificação única, URL e localização. Com o *DataCite*, foram inseridos assunto, colaborador, formato, tamanho, descrição, idioma, direitos e financiador (SILVELLO, 2018).

Figura 3 – Exemplo de referência de acordo com o *DataCite* (2016)



Fonte: elaborado pelas autoras (2019) com base no exemplo do *DataCite*.

Trabalhos posteriores a 2016 descreveram o *DataCite* como um padrão de citação que almeja seguir os princípios descritos em JDDCP. Buneman, Davidson e Frew (2016) e Silvello (2018) analisaram a aderência do crescente número de citações de dados ao padrão *DataCite*, identificando, também, suas deficiências, tais como a falta de metadados temporais de coleta dos dados, a dificuldade de atribuição de identificador persistente para subconjuntos de dados extraídos de bases de dados dinâmicas, bem como para agregações de conjuntos de dados distintos, a falta de forma automatizada para a criação de referência que descreva a procedência dos dados de forma consistente (SILVELLO, 2018).

Apesar de cada conjunto de dados ter um padrão de citação sugerido, os pesquisadores têm adotado diferentes formas para citar os conjuntos de dados. Para ser exato, 377 métodos foram encontrados para “citar diferentes versões de um único conjunto de dados” (BELTER, 2014, p. 7), causando múltiplos pontos de acesso para eles. Robinson-García, Jiménez-Contreras e Torres-Salinas (2016), com base em levantamento feito no DCI, afirmam que a citação de dados não está entre as práticas comuns na maioria dos campos de pesquisa, sendo pouco expressiva, exceto para as áreas de Engenharia e Tecnologia, que costumam citar mais os dados, e que operam de forma distinta das áreas de Ciências Sociais, Artes e Humanas, que citam mais os estudos relacionados aos dados. Esse resultado quanto ao consumo de dados ser mais frequente nas exatas é comum, tendo em vista, principalmente, as distintas abordagens metodológicas e necessidades das áreas.

Silvello (2018) defende a identificação única para o conjunto e subconjunto de dados com suporte para referência de agregação de conjuntos de dados, sugerindo uma hierarquia de identificadores: um *Persistent Identifier* (PID) (DOI, *handle*) ou um *Research Resource Identifier* (RRID) como raiz + *Universal Numerical Fingerprint* (UNF) (um número associado ao conteúdo semântico do dado que recupera o dado independentemente do seu formato) ou + consulta *Xquery* ou *SQL*.

Park e Wolfram (2019) investigaram a citação de *softwares* usados para pesquisa (*research software*). Sua pesquisa revelou que 99% dos *softwares* indexados no DCI pertencem a cinco repositórios: 63,21% registram os dados científicos no Zenodo (que tem ligação com o GitHub); 24,75% no *Comprehensive R Archive Network* (CRAN); 3,78% no nanoHub; 2,96% no ModelDB (neurociência); 2,69% no FigShare e 2,28% no ASCL (Astronomia). Os autores também analisaram quais os metadados mais incidentes

na referência de *software*: autor (99,2%), título (100%), descrição (99,5%), idioma (100%), categoria (100%), ano (100%) e tipo (92,2%) foram amplamente utilizados, entretanto, o DOI (49,4%) não. Para os autores, o uso do DOI não impactou o aumento de citações, sendo mais usado por FigShare, nanoHUB e Zenodo; além disso, os outros repositórios mantinham a URL. Outra revelação da referida pesquisa foi a falta de identificador padronizado para pesquisadores, como o uso do ORCID ou outro ID. Concluíram que a taxa de citação de *software* é pequena, sendo um dos possíveis motivos a falta de padrão de citação, feita muitas vezes no corpo do texto, local que o DCI não detecta. Sendo assim, os pesquisadores não puderam afirmar se ocorre reuso, pois isso pode se dar sem o uso de citação correta.

Referente à citação automática, Buneman, Davidson e Frew (2016) propõem uma metodologia baseada em “*views*” de bases de dados e “*queries*” para criar automaticamente a citação de dados. Os autores demonstram como seria na prática em duas bases: GtoPdb (*Gide to Pharmacology*, com hierarquias de famílias de medicamentos) e MODIS (fotos de satélites). Segundo os autores, apesar de ser recomendável o uso de identificadores persistentes, isso não garante a imutabilidade (*fixity*), ou seja, que o dado citado não mude. Para esses autores, o formato *DataCite* permite a parametrização de coordenadas geográficas, porém, não permite de tempo, o que dificulta adequá-lo para a base MODIS.

Conforme revelado pela pesquisa de Silvello (2018), os sistemas em que há suporte para criação automática de “texto de citação” são específicos de áreas de conhecimento e não de propósito geral. No caso desses últimos, o “texto de citação” é manual, o que tende a causar inconsistências nas citações, pois o usuário pode não ter o conhecimento (do ambiente e mesmo técnico) para criá-lo.

Atualmente, os principais repositórios de dados de propósito geral (Dataverse, FigShare, Dryad, Mendeley Data, Zenodo, DataHub, DANS, EUDat) não suportam granularidade variada (citação de conjunto, subconjunto ou agregação de conjuntos de dados) nem criam “textos de citação” apropriados e informativos de forma automática, sendo necessário o desenvolvimento de ferramentas fáceis de usar e que tenham visualização gráfica atrativa (SILVELLO, 2018).

Alguns estudos mostram estratégias para lidar com os conjuntos de dados dinâmicos. Para Callaghan, Lowry e Walton (2012, p. 9), “não há garantia para o usuário da citação de que o conjunto de dados recuperado em uma data será o mesmo de quando a citação foi escrita em algum momento anterior”, visão essa compartilhada por

Pröll e Rauber (2015, p. 26), para quem “a questão sobre como permitir a citação de um subconjunto de dados, estáticos e, principalmente, dinâmicos, de forma escalável e precisa permanece como uma tarefa não trivial”.

Segundo esses autores, a referência pode ser afetada dependendo da estabilidade do conjunto de dados. Para dados estáticos, é possível quebrar o conjunto de dados em partes menores, atribuir um DOI a cada uma dessas partes e, no futuro, quando o conjunto de dados for finalizado, atribuir um único DOI à família dos dados. Quando os dados são dinâmicos e recebem atualizações contínuas, bem como, por vezes, novas inserções de dados, é mais apropriado registrar *fixed snapshots*, que servem para deixar estável o conjunto de dados, para, com base nisso, “armazená-lo em outro local no repositório e, em seguida, atribuir o DOI a esse instantâneo específico” (CALLAGHAN; LOWRY; WALTON, 2012, p. 10).

Para Mathiak e Boland (2015), a atribuição do DOI também é importante, apesar de terem observado comportamentos distintos relacionados a isso, com o registro de até três DOI para um mesmo conjunto de dados em um determinado repositório: um para os dados coletados pessoalmente por uma assistente, um para os dados coletados automaticamente por máquinas e um terceiro para ambos os conjuntos de dados integrados. No entanto, alguns repositórios registram um DOI para cada conjunto de dados incremental.

Outra opção seria uma estratégia centrada em consulta (*query center view*) para a citação de dados dinâmicos, que pode ser aplicada em dados versionados e datados, em que qualquer inserção, atualização ou exclusão de dados é registrada. É atribuído um identificador persistente para cada consulta (*query*), que guarda a data em que esta foi executada. Dessa forma, é possível recuperar o estado do conjunto de dados no momento em que a consulta citada foi executada para o respectivo trabalho (PRÖL & RAUBER, 2015).

Na área de Física, o consórcio *Virtual Atomic and Molecular Data Centre*² (VAMDC Consortium) considera que a ciência intensiva de dados tem ainda como âncora a citação e, em virtude disso, estudou diferentes maneiras de citar os dados com base na

² “Compartilham entre Institutos e Instituições de Pesquisa uma estrutura técnica e política comum para a distribuição e a curadoria de dados atômicos e moleculares”, voltada para astrofísica, física atmosférica, física de plasma, biofísica. Países parceiros dessa iniciativa são Austrália, Áustria, França, Alemanha, Índia, Itália, Coreia, Rússia, Sérvia, África do Sul, Suécia, Reino Unido, EUA, Venezuela. No *site*, designa um padrão de citação adotado indicando inclusive como o consórcio deve ser citado (VAMDC Consortium, 2019, p. 1).

orientação RDA³ para citação de dados dinâmicos (ZWÖLF et al., 2019). Utilizando o XML como linguagem, criou um modelo de extração de dados para citação e conversão gráfica por meio de um método não ambíguo para rastrear e documentar os fluxos das reutilizações dos dados (ZWÖLF; MOREAU; DUBERNET, 2016). O RDA propôs que a citação seja centrada no controle de versão, no registro de data e hora dos dados, assim como na consulta⁴; para isso, foi necessária a configuração de um *Query Store* (ZWÖLF et al., 2019). O RDA Scholix é um *framework*⁵ que tem como objetivo aumentar a “interoperabilidade e permitir um ecossistema de informações aberto” (ZWÖLF et al., 2019, p. 2). Nesse caso, os autores propõem que sejam consideradas para os dados dinâmicos duas versões de citação que incorporem o controle de versionamento de dados e registros de data e hora: granulação grossa (possibilita modificar qualquer dado público em um determinado nó de dados, induzindo ao incremento da versão do nó, ou seja, um mecanismo que informe apenas que houve alteração); granulação fina (significa que as informações do “versionamento indicam quais dados foram alterados entre duas versões de nó de dados diferentes”) (ZWÖLF et al., 2019, p. 2).

Os autores integraram o serviço *Query-Store* por intermédio da *REST API* do *Zenodo*, considerando apenas a granulação grossa. Concluíram que, com a disponibilização do serviço de *Query-Store* no *Zenodo*, houve um aumento nas visualizações de consultas já cadastradas e na criação de novas consultas, sendo esse monitoramento possível porque a cada consulta nova e única é possível atribuir um DOI (ZWÖLF et al., 2019).

De acordo com a presente revisão, apesar das divergências e necessidades distintas das áreas, a tendência é que adotem um único padrão, levando em conta suas peculiaridades e a granularidade do conjunto de dados. Esse padrão teria metadados obrigatórios e opcionais, tendo em vista que alguns dados exigem recursos, por exemplo, uma especificação de temporalidade maior (ZWÖLF et al., 2019). A adoção de padrão único também facilitaria a descoberta e recuperação automática feita por meio de máquinas.

³ A Research Data Alliance (RDA) é uma organização internacional que reúne comunidades de pesquisadores interessados em estudar e promover o compartilhamento aberto de dados científicos.

⁴ Consulta é definida como “qualquer mecanismo de processamento usado para extrair dados” (ZWÖLF et al., 2019).

⁵ Conjunto de funções de *software* que têm um objetivo em comum.

3.3 Análises métricas para citação de dados

Esta seção trata das evidências de pesquisas a respeito das métricas para dados científicos, ou seja, são realizados os mesmos tipos de estudos métricos de informação em conjuntos de dados?

O estudo mais antigo abrangido pela presente pesquisa (SIEBER; TRUMBO, 1995) utiliza uma bibliografia do *General Social Survey* (GSS, conjunto de dados em Ciências Sociais) disponibilizada pelo *National Opinion Research Center* (NORC) para identificar 198 artigos publicados entre 1976 e 1988 que reutilizaram dados compartilhados. As citações do conjunto original de dados foram procuradas manualmente nas publicações e, apesar de estudos métricos tradicionais considerarem apenas a seção de referências, Sieber e Trumbo (1995) analisaram todo o corpo do texto em busca de citações ou menções aos dados originais, ainda que o espaço destinado às referências recebesse maior relevância. Constataram que, nessa seção, somente 13% mencionaram o autor principal dos dados e 19%, o nome da pesquisa, sendo que 60% não declararam o nome do pesquisador e 9% não mencionaram o nome da pesquisa em parte alguma do texto.

Na área da Saúde, o estudo de 2007 realizado por Piwowar, Day e Fridsma utilizou o método de regressão multivariada para contagem de citações, juntamente com a correlação dos dados. A seleção do *corpus* de 85 ensaios foi realizada por meio de revisão sistemática publicada por terceiros, abarcando os anos de 1999 a 2003, dos quais 41 ensaios tinham seus dados disponíveis publicamente em diferentes *sites*: do laboratório, de editores ou em bancos de dados públicos (no banco de dados *Stanford Microarray*, em *Gene Expression Omnibus*, em *ArrayExpress*, Portal de Dados *GeneExpression* do NCI). O número de citações recebeu transformação logarítmica e, com base na taxa de citação, foi correlacionado com o fator de impacto da revista que publicou cada estudo, a data de publicação e o país dos autores, de modo que esses fatores fossem incluídos como covariáveis. Os autores concluíram que tornar os dados públicos está significativamente relacionado a um aumento de 69% (com um intervalo de confiança de 95% o aumento pode estar entre 18% e 143%) nas citações dos artigos (PIWOWAR; DAY; FRIDSMA, 2007).

Em 2013, seis anos após a pesquisa anterior em que Piwowar, Day e Fridsma (2007) trouxeram evidências do aumento de citações por conta de disponibilização dos dados, Piwowar e Vision (2013) repetiram a mesma metodologia e obtiveram um resultado distinto no que se refere ao crescimento de citações. A última pesquisa, com

maior controle e padronização dos dados, obteve um resultado de 9% no aumento de citações para artigos que publicaram seus dados. Os autores adicionaram covariáveis diferentes do estudo anterior: *status* de acesso aberto, número de autores, primeiro e último histórico de publicação do autor e tópico do estudo. Com o foco na reutilização dos dados por uma terceira parte, excluíram citações de conjuntos de dados cujo sobrenome do autor fosse o mesmo do artigo analisado.

Com base em métodos bibliométricos, Belter (2014) buscou gerar indicadores de citação e medir o impacto de três conjuntos de dados disponibilizados em acesso aberto, com curadoria e arquivamento pelo *National Oceanic and Atmospheric Administration* (NOAA) da *National Oceanographic Data Center* (NODC). O resultado foi que grande parte das citações encontradas não ocorriam em campos legíveis na base de dados consultada, WoS (título, resumo, agradecimentos, referências), obtendo contagens de citação maiores no *Google Scholar* e *sites* das editoras que analisam o texto completo.

Em 2015, Henderson e Kotz analisaram as citações de conjuntos de dados do *Community Resource for Archiving Wireless Data At Dartmouth* (CRAWDAD) que, apesar de solicitar a seus usuários a utilização do repositório e do gerenciador de referências CiteULike, poucos de fato o fazem. Isso foi comprovado por buscas pelo termo “CRAWDAD” nas bases de dados *ACM Digital Library*, *Google Scholar*, *IEEE Xplore*, *ScienceDirect* e *Scopus* para localizar as publicações que citavam algum de seus conjuntos de dados. Os autores avaliaram os 1.295 resultados finais com critérios baseados nos *Data Citation Principles* do FORCE11: crédito e atribuição, identificação única, acesso e persistência. Segundo esses autores, dos 1.281 artigos que puderam ser acessados, 88% cumpriam os critérios mínimos estabelecidos. Entre os que não cumpriram, destaca-se a citação dos artigos originais dos dados compartilhados e não do conjunto de dados diretamente.

Mathiak e Bolland (2015) consideraram que, nas Ciências Sociais, os levantamentos amostrais (*surveys*) foram predominantes entre os dados científicos obtidos e que seus resultados podem variar bastante em relação a como as entrevistas são executadas. Portanto, os identificadores persistentes são atribuídos às partes individuais menores do conjunto de dados. Um exemplo seria designar três DOI distintos às entrevistas cujas respostas tenham sido coletadas em suportes diferentes – papel, computador e misto – ou atribuir DOI pelo ano em que foram obtidas. Porém, essa divisão não é padronizada e a decisão é tomada para cada estudo, dificultando, assim, a

identificação do conjunto de dados específico por meio das referências (MATHIAK; BOLAND, 2015).

Com o avanço dos estudos métricos relacionados à citação de dados, foi criado um algoritmo que encontra automaticamente as citações de conjuntos de dados; porém, permanece o problema de relacionar a citação do DOI correspondente (MATHIAK; BOLAND, 2015). As citações costumam não detalhar informações que possibilitam diferenciar as versões dos conjuntos de dados (variações de ano, amostras, países etc.). Para relacioná-las a seus respectivos conjuntos de dados e os próprios conjuntos entre si, foi proposto por Mathiak e Boland (2015) a utilização de vocabulário controlado por meio de um sistema de metadados, o que também facilitaria a recuperação dos dados.

Na área Interdisciplinar, em 2016, foi utilizado o DCI para investigar a citação de dados por meio da análise de distribuição de citações, abrangendo os registros de 1951 a 2013. O estudo identificou quais áreas estão citando dados científicos (81% Ciências, 18% Ciências Sociais, Artes e Humanas 2%, 0,1% Engenharia e Tecnologia) e um comportamento de citação de dados de acordo com a sua fonte de origem: os mais citados (294.051) são conjunto de dados, os estudos dos dados (106.895) em segundo lugar e, por último, os (3.265) repositórios de dados (ROBINSON-GARCÍA; JIMÉNEZ-CONTRERAS; TORRES-SALINAS, 2016). Segundo os autores, o estudo cobriu todos os registros de dados indexados no DCI, criando um banco de dados relacional para processar os dados. Eles frisaram que as áreas das Engenharias e Tecnologias podem estar mal representadas no DCI porque possuem outros recursos para o armazenamento dos dados e para outras práticas de citação. Também foi realizado um levantamento dos assuntos estudados, no entanto, os autores reconheceram que ocorreu alto grau de distorção, que pode ter sido ocasionado pela falta de padronização dos metadados.

O DCI foi também utilizado por Onyancha (2016) para verificar o compartilhamento de dados científicos na África Subsaariana, na qual foi aplicada uma estratégia de busca contendo todos os 50 países da região. Os resultados foram analisados através da própria ferramenta do DCI, permitindo a análise de autoria, países/territórios de origem, tipos de documentos, editoras, coautoria, instituições, línguas, títulos, assuntos, categorias da WoS e anos das publicações. As citações dos conjuntos de dados foram comparadas em análise correlacional com as citações de artigos na mesma região; verificou-se que os índices de citação dos conjuntos de dados foram muito inferiores às publicações tradicionais. É possível que esse resultado seja observado em razão do método da pesquisa considerar apenas as citações formais. Posto que os padrões de

citação de dados são uma realidade recente, as normas de referência ainda não são dominadas pelos autores e muitas citações são perdidas.

Uma das mais importantes iniciativas acerca de dados científicos na área de Física de Altas Energias é o INSPIRE-HEP, uma base de dados que passou a agregar também os dados científicos, conectando-se a fontes como Dataverse, FigShare e HEPData (HERTERICH; DALLMEIER-TIESSEN, 2016). O maior diferencial do INSPIRE é seu sistema de contagem de citações que, utilizando mineração de texto e análise bibliográfica manual, engloba a citação de dados e a direciona às publicações que os citam, bem como aos perfis de seus autores. Herterich, Dallmeier-Tiessen (2016) questionam a utilização de métricas de citação em conjuntos de dados ou coleções maiores de dados.

A metodologia da altmetria para citação de dados foi utilizada por Peters et al. (2016) com o objetivo de verificar, por meio do DCI, se a citação de dados impactava as plataformas de mídias sociais. Considerando apenas os itens que receberam mais de duas citações, para evitar casos de autocitação, uma vez que estes não são discriminados pelo DCI, foram aplicadas ferramentas como o ImpactStory, Altmetric.com e PlumX (a que tem melhor cobertura) às citações que continham DOI. Foi revelado que 85% das pesquisas não foram mencionadas nas redes. Logo, concluíram que há baixa relevância altmétrica relacionada aos dados científicos.

Em 2017, as características das áreas de Genética e Hereditariedade foram examinadas utilizando-se o *Data Citation Index* (DCI) da WoS para verificar citações registradas (PARK; WOLFRAM, 2017). Park e Wolfram (2017) optaram pela análise centrada no autor citante para remediar a autocitação de dados e encontrar citações implícitas dentro das 148 publicações amostradas. Como objetivo mais geral, os autores procuravam identificar fatores que influenciavam o compartilhamento e a reutilização de dados. As citações foram buscadas em referências, texto principal, informações suplementares, reconhecimentos, informações de financiamento, informações do autor e recursos da *web*. Eles constataram que a taxa de autocitação para dados (8%) é bem maior que para artigos (1,2%) e que a quantidade total de citações referentes ao reuso dos dados (66) é bem menor que a referente ao compartilhamento dos dados (316).

Zhao, Yan e Li (2018) aplicaram pressupostos da análise de conteúdo como instrumento de pesquisa. Para isso, criaram um esquema de decodificação textual que considera cada frase como uma unidade de análise, avaliando tanto as citações formais como menções aos conjuntos de dados. Dos 600 artigos inicialmente selecionados na

PLoS One, 52% utilizaram conjuntos de dados em suas pesquisas, dos quais 74% foram provenientes de autocitação, ou seja, os autores reutilizaram os próprios dados publicados; 60% dos artigos que utilizavam dados forneceram a URL para o conjunto de dados, enquanto 24% mencionaram apenas o nome deste; somente 9% atribuíram identificadores. A ferramenta criada pelos autores ainda permitiu comparar a utilização de citação dos conjuntos de dados nas disciplinas abrangidas e identificar as áreas que mais utilizavam esses dados, sendo as principais Ciências da Saúde (90%) e Ciências Sociais (66%).

Na área de Ciência da Informação, Park, You e Wolfram (2018), com base em uma experiência de estudo anterior (PARK; WOLFRAM, 2017), criaram um método que automatiza o processo de extração do texto com validação humana, com o objetivo de indicar as ocorrências de compartilhamento e reutilização dos dados, inserindo no sistema os termos com probabilidade de revelar sua existência. O estudo usou o DCI, sendo que a amostra do conjunto de dados foi das áreas de Ciências Biológicas e Biomedicina. Os DOI encontrados nas citações foram convertidos por meio de uma API no padrão de identificadores do PUBMED *Central*. Para analisar os resultados gerados pela API, usaram o Python SOAP (*Simple Object Access Protocol*). O resultado evidenciou que o motivo pelo qual não houve a contabilização de citações de conjuntos de dados deu-se em decorrência de que citações informais de dados não são consideradas em fontes, como no próprio DCI.

A crítica recorrente à falta de padronização pode ser verificada por meio dos erros mais comuns ao citar os dados, como mencioná-los fora da seção de referências ou citar instâncias correlatas, como o próprio repositório de dados ou o artigo original que os utilizou. Portanto, é perceptível que, embora seja possível realizar estudos métricos que identifiquem as citações manualmente (SIEBER; TRUMBO, 1995; BELTER, 2014; HENDERSON; KOTZ, 2015; PARK; WOLFRAM, 2017; PIWOWAR; DAY; FRIDSMA, 2007), muito ainda precisa ser feito para permitir análises métricas em larga escala ou automatizadas (PARK; YOU; WOLFRAM, 2018).

4 DISCUSSÃO

Nesta revisão, elaborada com trabalhos distribuídos entre 1995 e 2019, houve o consenso de que o padrão de citação dos produtos tradicionais da pesquisa não atende às necessidades de citação de dados das diferentes áreas.

Observou-se a tendência das áreas para buscar uma definição de um padrão de citação que atenda às demandas de todos os tipos de dados, incrementando padrões existentes com novas informações, desde que seja possível incorporar dados pertinentes à tipologia e estabilidade do conjunto de dados. O trabalho desenvolvido pelo FORCE11 na elaboração da *Joint Declaration of Data Citation Principles* (JDDCP) foi fundamental para a evolução dos padrões de citações de dados, pois definiu aquilo que era imprescindível para ser atendido por um padrão, independentemente da área. Dessa forma, foi possível traduzir esses princípios em elementos reais conforme a área. Nesse cenário, despontam o formato *DataCite* de citação, com a versão 4.3 do seu esquema de metadados lançada em agosto de 2019⁶, e o DOI como identificador único.

Há um esforço da comunidade científica para viabilizar a citação de dados por meio da automatização, assim como aplicações para sua análise, desde que seja ponderado o uso desses resultados quantitativos para recompensar os autores. A análise de citação de dados científicos pode ser considerada como um novo campo de estudo.

A principal dificuldade nas análises de citação de dados científicos (Quadro 3) é justamente a sua falta de padronização, a qual limita os estudos automáticos de citação, ou seja, mesmo com a possibilidade de contagem de citações automáticas e índices de citações de dados, os estudos métricos mais abrangentes ainda precisam recorrer à busca e contabilização manual de citações (menções informais), bem como aos conjuntos de dados no corpo do texto, em virtude do desconhecimento dos autores a respeito da necessidade de padronização da citação de dados científicos.

Quadro 3 – Métodos de coleta de dados dos artigos

Ano	Autores	Método
1995	Sieber; Trumbo	Análise de um <i>dataset</i> para verificar menções em artigos e padrões de citações; questionário para autores.
2007	Piwowar; Day; Fridsma	Procura <i>datasets</i> associados a um grupo de artigos verificando o corpo dos textos e repositórios de dados.
2014	Belter	Análise de <i>datasets</i> para gerar contagem de citações ou menções por meio de bases de dados.
2015	Henderson; Kotz	Análise de <i>datasets</i> para gerar contagem de citações ou menções por meio de bases de dados e verificação de uso dos princípios de citação do Force11.
2015	Mathiak; Boland	Não possui seção de metodologia.
2016	Herterich; Dallmeier-Tiesse	Não possui seção de metodologia.
2016	Onyancha	Analisa <i>datasets</i> de países da África Subsaariana disponíveis no DCI.
2016	Peters et al.	Utiliza DCI para recuperar dados científicos citados ao menos duas vezes e relaciona-os com índices altmétricos.

⁶ DataCite Schema – <https://schema.datacite.org/>

Ano	Autores	Método
2016	Robinson-García; Jiménez-Contreras; Torres-Salinas	Analisa a distribuição de citações no DCI por área e repositório.
2017	Park; Wolfram	Analisa manualmente as citações dos conjuntos de dados do DCI relacionadas à WoS, verificando referências, texto principal, informações suplementares, reconhecimentos, informações de financiamento, informações do autor e recursos da <i>web</i> .
2018	Park; You; Wolfram	Coletam registros de dados no DCI na área de genética e hereditariedade. Identifica citações formais e informais.
2018	Zhao; Yan; Li	Realizam análise de conteúdo em artigos de acesso aberto da PLoS One para identificar padrões de menções e citações de dados.

Fonte: Dados da pesquisa. Adaptado pelas autoras (2019).

Cabe salientar que, apesar da importância de dar credibilidade ao autor, nenhum dos artigos encontrados apontou a relevância das licenças apropriadas para o compartilhamento e reuso dos dados, por exemplo, as licenças *Creative Commons* e *Open Data Commons* (SILVA, 2019).

Outro ponto a frisar é que, apesar da relevância das plataformas de dados bem estruturadas e dos planos de gestão de dados, pouco se tem discutido a respeito do desenvolvimento de programas de preservação em longo prazo. Quando um documento perde o enlaçamento com o conjunto de dados que usa para fundamentar suas afirmações e conclusões, ocorre uma reação em cadeia tanto para o autor original quanto para as reutilizações, o que culmina em perda do valor da pesquisa (PARK; YOU; WOLFRAM, 2018; CALLAGHAN, 2014). Mesmo quem cita o dado precisa se preocupar com sua preservação em longo prazo, tendo em vista que sua pesquisa poderá ser afetada. Significa ainda que os pesquisadores, instituições de fomento, institutos e universidades precisam prever nos projetos de pesquisa os valores relacionados à preservação e manutenção dos conjuntos de dados, considerando o custo médio de 1,5% do total das despesas de pesquisas (PARK; YOU; WOLFRAM, 2018). Essa é uma responsabilidade que não pode ser dos autores. Deve ser confiada às instituições públicas governamentais ou comerciais, com permissão de embargo de acesso, se for o caso (MAYERNIK, 2012).

De modo geral, a revisão aponta para um pequeno número de artigos relevantes e apenas em língua inglesa. Isso demonstra que o assunto ainda precisa ser explorado em pesquisas científicas, pois trata-se de uma área em expansão e de fundamental importância para a comunicação científica.

4 CONCLUSÃO

Os dados científicos sempre existiram no contexto de uma pesquisa, no entanto, com os recursos da internet e as necessidades de consistência, integridade, reprodutibilidade e replicabilidade de estudos, faz-se necessário incorporar o universo dos dados na rotina da produção científica.

Em virtude de sua atual relevância, a citação de dados tem tomado dimensões que afetam todos os envolvidos no processo da produção científica, seja uma pesquisa publicada em canais formais ou não. Isso demanda esforços do pesquisador, da instituição afiliada, das agências de fomento, dos repositórios de dados, das equipes editoriais de periódicos tradicionais e de dados.

Para viabilizar os estudos de análises de citação de dados científicos são necessárias ações integradas, com o estabelecimento de políticas de acordo com as necessidades das áreas de conhecimento. Cada segmento possui necessidades específicas que precisam ser consideradas de tal modo que os metadados ainda possam interoperar com outros sistemas e outras áreas.

Um aspecto proeminente, mas pouco discutido, foi o aumento da percepção de valor da pesquisa por meio da disponibilização dos dados científicos. Isso porque, quando os conjuntos de dados são compartilhados, outras áreas podem se beneficiar deles e utilizá-los com perspectivas diferentes da pesquisa original. Isso pode ser uma vantagem política e científica, na qual a colaboração entre países poderá se fortalecer. Nos países em desenvolvimento, por exemplo, em que há pouco recurso para investir em equipamentos sofisticados, os pesquisadores poderão colaborar com o aprofundamento teórico baseado nos conjuntos de dados científicos compartilhados, caso sua disponibilização se torne uma realidade consistente.

Consideramos a promoção do avanço da ciência uma das maiores vantagens da disponibilização de dados, visto que esta pode ser ofuscada pelas dificuldades técnicas de pesquisadores ao publicarem seus dados, assim como por outros fatores que podem gerar incredibilidade ou atrapalhar sua vantagem competitiva quanto à originalidade de suas análises de pesquisa. Isso pode ser considerado não apenas uma mudança cultural quanto à disponibilização de dados, mas também de pertencimento e de bem comum, que caracteriza o conhecimento humano. Se o conhecimento é construído considerando pesquisas anteriores, logo, caracteriza-se como uma colaboração, pois o conhecimento não é individual, firma-se no coletivo. Assim, as estruturas acadêmicas e de pesquisa

precisarão se adaptar a esse novo contexto, mudando principalmente sua cultura de produção científica.

REFERÊNCIAS

ALTMAN, M. et al. An introduction to the joint principles for data citation. **Bulletin of the American Society for Information Science and Technology**, v. 41, n. 3, p. 43–45, fev. 2015. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.2015.1720410313>. Acesso em: 19 ago. 2019.

ALTMAN, M.; MERCÈ CROSAS. The Evolution of Data Citation: From Principles to Implementation. **IASSIST Quarterly**, v. 37, n. 1, p. 62, 14 jun. 2013. Disponível em: https://iassistdata.org/sites/default/files/iqvol371_4_altman.pdf. Acesso em: 19 ago. 2019.

BELTER, Christopher W. Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. **PLoS ONE**, [s. l.], v. 9, n. 3, p. e92590, 2014. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0092590>. Acesso em: 22 jun. 2019.

BUNEMAN, P.; DAVIDSON, S.; FREW, J. Why data citation is a computational problem. **Communications of the ACM**, v. 59, n. 9, p. 50–57, 24 ago. 2016. Disponível em: <https://dl.acm.org/citation.cfm?id=2893181>. Acesso em: 19 ago. 2019.

CALLAGHAN, Sarah; LOWRY, Roy; WALTON, David. Data Citation and Publication by NERC's Environmental Data Centres. **Ariadne: Web Magazine for Information Professionals**, n. 68, 2012. Disponível em: <http://www.ariadne.ac.uk/issue/68/callaghan-et-al/>. Acesso em: 24 jun. 2019.

CALLAGHAN, Sarah. Preserving the integrity of the scientific record: data citation and linking. **Learned Publishing**, [s. l.], v. 27, n. 5, p. 15–24, 2014. Disponível em: <http://doi.wiley.com/10.1087/20140504>. Acesso em: 24 jun. 2019.

CODATA Task Group on Data Citation Standards and Practices, Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, **Data Science Journal**, 2013, Vol. 12, <https://doi.org/10.2481/dsj.OSOM13-043>

COUSIJN, Helena et al. A data citation roadmap for scientific publishers. **Scientific Data**, v. 5, p. 180-259, 2018. Disponível em: <http://www.nature.com/articles/sdata2018259>. Acesso em: 22 jun. 2019.

DATA CITE. **DataCite – Find, access and Reuse Data**. 2019. Disponível em: <https://www.datacite.org/>. Acesso em: 22 jun. 2019.

DRYAD Digital Repository. **Dryad Digital Repository (DRYAD)**. 2019. Disponível em: <http://datadryad.org/>. Acesso em: 30 jul. 2019.

FIGSHARE. FigShare Credit for all Research. 2019. Disponível em: <https://figshare.com/>. Acesso em: 30 jul. 2019.

FORCE11, Data Citation Synthesis Group: Joint declaration of data citation principles, Martone, M. (ed) San Diego, CA. 2014, Disponível em: <https://doi.org/10.25490/a97f-egykh>. Acesso em: 14 ago. 2019.

FORCE, M. M.; ROBINSON, N. J. Encouraging data citation and discovery with the Data Citation Index. **Journal of Computer-Aided Molecular Design**, v. 28, n. 10, p. 1043–1048, 1 out. 2014. Disponível em: <http://link.springer.com/10.1007/s10822-014-9768-5>. Acesso em: 22 jun. 2019.

HERTERICH, P., DALLMEIER-TIESSEN, S. Data Citation Services in the High-Energy Physics Community. 2016, v. 22, n. 1/2, 1. <https://doi.org/10.1045/january2016-herterich>

HENDERSON, Tristan; KOTZ, David. Data Citation Practices in the CRAWDAD Wireless Network Data Archive. **D-Lib Magazine**, v. 21, n. 1/2, 2015. Disponível em: <http://www.dlib.org/dlib/january15/henderson/01henderson.html>.

JBIR REVIEWER'S MANUAL. Disponível em: <https://wiki.joannabriggs.org/m/view-rendered-page.action?abstractPageId=3178510>. Acesso em: 19 jan. 2020.

LI, Kai; CHEN, Pei-Ying. The narrative structure as a citation context in data papers: A preliminary analysis of Scientific Data. **Proceedings of the Association for Information Science and Technology**, v. 55, n. 1, p. 856–858, 2018. Disponível em: <http://doi.wiley.com/10.1002/pa2.2018.14505501147>. Acesso em: 24 jun. 2019.

MAYERNIK, Matthew S. Data citation initiatives and issues. **Bulletin of the American Society for Information Science and Technology**, [s. l.], v. 38, n. 5, p. 23–28, 2012. Disponível em: <http://doi.wiley.com/10.1002/bult.2012.1720380508>. Acesso em: 24 jun. 2019.

MATHIAK, B., & BOLAND, K. Challenges in Matching Dataset Citation Strings to Datasets in Social Science. **D-Lib Magazine**, 21(1/2), 2015. <https://doi.org/10.1045/january2015-mathiak>

MOONEY, Hailey. A Practical Approach to Data Citation: The Special Interest Group on Data Citation and Development of the Quick Guide to Data Citation. **IASSIST Quarterly**, [s. l.], v. 37, n. 1, p. 71, 2013. Disponível em: <https://iassistquarterly.com/index.php/iassist/article/view/240>. Acesso em: ago. 2019.

MOONEY, H., & NEWTON, M. The Anatomy of a Data Citation: Discovery, Reuse, and Credit. **Journal of Librarianship and Scholarly Communication**, eP1035, 2012. <https://doi.org/10.7710/2162-3309.1035>. Acesso em: ago. 2019.

NOVACESCU, Jenny et al. A Model for Data Citation in Astronomical Research Using Digital Object Identifiers (DOIs). **The Astrophysical Journal Supplement Series**, [s. l.], v. 236, n. 1, p. 20, 2018. Disponível em: <http://stacks.iop.org/0067-0049/236/i=1/a=20?key=crossref.b275284de5e1758e81449761f021c029>. Acesso em: 22 jun. 2019.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. OECD Principles and Guidelines for Access to Research Data from Public Funding. Paris: OECD,

2007. 24 p. Disponível em: <http://www.oecd.org/sti/inno/38500813.pdf>. Acesso em: 17 abr. 2020.

ONYANCHA, Omwoyo Bosire. Open Research Data in Sub-Saharan Africa: A Bibliometric Study Using the Data Citation Index. **Publishing Research Quarterly**, [s. l.], v. 32, n. 3, p. 227–246, 2016. Disponível em: <http://link.springer.com/10.1007/s12109-016-9463-6>. Acesso em: 14 maio 2019.

PARK, Hyunjung; WOLFRAM, Dietmar. An examination of research data sharing and re-use: implications for data citation practice. **Scientometrics**, Dordrecht, v. 111, n. 1, p. 443–461, 2017. Disponível em: <http://link.springer.com/10.1007/s11192-017-2240-2>. Acesso em: 25 abr. 2019.

PARK, H.; WOLFRAM, D. Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse. **Journal of Informetrics**, v. 13, n. 2, p. 574–582, maio 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157718302372>. Acesso em: 19 ago. 2019.

PARK, Hyunjung; YOU, Sukjin; WOLFRAM, Dietmar. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. **Journal of the Association for Information Science and Technology**, v. 69, n. 11, p. 1346–1354, 2018. Disponível em: <http://doi.wiley.com/10.1002/asi.24049>. Acesso em: 22 jun. 2019.

PAVLECH, Laura L. Data Citation Index. **Journal of the Medical Library Association**, v. 104, n. 1, p. 88–90, jan. 2016. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722655/pdf/mlab-104-01-88.pdf>. Acesso em: 19 ago. 2019.

PETERS, Isabella et al. Research data explored: an extended analysis of citations and altmetrics. **Scientometrics**, [s. l.], v. 107, n. 2, p. 723–744, 2016. Disponível em: <http://link.springer.com/10.1007/s11192-016-1887-4>. Acesso em: 24 jun. 2019.

PIWOWAR, Heather A.; DAY, Roger S.; FRIDSMA, Douglas B. Sharing Detailed Research Data Is Associated with Increased Citation Rate. **PLoS ONE**, v. 2, n. 3, p. e308, 2007. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0000308>. Acesso em: 14 maio 2019.

PIWOWAR, Heather A.; VISION, Todd J. Data reuse and the open data citation advantage. **PeerJ**, [s. l.], v. 1, p. e175, 2013. Disponível em: <https://peerj.com/articles/175>. Acesso em: 22 jun. 2019.

PRÖLL, S.; RAUBER, A. Asking the Right Questions - Query-Based Data Citation to Precisely Identify Subsets of Data. **{ERCIM} News**, n. 100, 2015. Disponível em: <https://ercim-news.ercim.eu/en100/special/asking-the-right-questions-query-based-data-citation-to-precisely-identify-subsets-of-data>. Acesso em: 19 ago. 2019.

ROBINSON-GARCÍA, N.; JIMÉNEZ-CONTRERAS, E.; TORRES-SALINAS, D. Analyzing data citation practices using the data citation index. **Journal of the Association for**

Information Science and Technology, v. 67, n. 12, p. 2964–2975, dez. 2016. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.23529>. Acesso em 19 ago. 2019.

SAYÃO, Luís Fernando; SALES, Luana Farias. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS**, v. 8, n. 2, p. 76-92, jun. 2014. DOI:10.3395/reciis.v8i2.934.pt. Disponível em: <http://www.reciis.icict.fiocruz.br/index.php/reciis/article/download/611/1252>. Acesso em: 09 ago. 2019.

SIEBER, J. E.; TRUMBO, B. E. (Not) giving credit where credit is due: Citation of data sets. **Science and Engineering Ethics**, v. 1, n. 1, p. 11–20, mar. 1995. Disponível em: <https://link.springer.com/article/10.1007/BF02628694>. Acesso em: 09 ago. 2019.

SILVA, F. C. C. **Gestão de Dados Científicos**. 1. ed. Rio de Janeiro: Interciência, 2019.

SILVELLO, Gianmaria. Theory and practice of data citation. **Journal of the Association for Information Science and Technology**, v. 69, n. 1, p. 6–20, 2018. Disponível em: <https://search.proquest.com/docview/1977750042?accountid=146814>. Acesso em: 09 ago. 2019.

SIMONS, N.; VISSER, K.; SEARLE, S. Growing Institutional Support for Data Citation: Results of a Partnership Between Griffith University and the Australian National Data Service. **D-Lib Magazine**, v. 19, nov. 2013. Disponível em: <http://www.dlib.org/dlib/november13/simons/11simons.html>. Acesso em: 19 ago. 2019.

WILKINSON et. al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, Nature, mar. 2016. Disponível em: <https://www.nature.com/articles/sdata201618.pdf>. Acesso em: 09 ago. 2019.

VAMDC CONSORTIUM. **Presentation**. 2019. Disponível em: <http://www.vamdc.org/structure/presentation/>. Acesso em: 24 jun. 2019.

ZHAO, Mengnan; YAN, Erjia; LI, Kai. Data set mentions and citations: A content analysis of full-text publications. **Journal of the Association for Information Science and Technology**, [s. l.], v. 69, n. 1, p. 32–46, 2018. Disponível em: <http://doi.wiley.com/10.1002/asi.23919>. Acesso em: 24 jun. 2019.

ZENODO. Zenodo Research Shared. 2019. Disponível em: <https://zenodo.org/>. Acesso em: 30 jul. 2019.

ZWÖLF, Carlo Maria et al. Implementing in the VAMDC the New Paradigms for Data Citation from the Research Data Alliance. **Data Science Journal**, [s. l.], v. 18, 2019. Disponível em: <http://datascience.codata.org/articles/10.5334/dsj-2019-004/>.

ZWÖLF, Carlo Maria; MOREAU, Nicolas; DUBERNET, Marie-Lise. New model for datasets citation and extraction reproducibility in VAMDC. **Journal of Molecular Spectroscopy**, [s. l.], v. 327, p. 122–137, 2016. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0022285216300613>. Acesso em: 24 jun. 2019.

NOTAS

AGRADECIMENTOS

Não se aplica.

CONTRIBUIÇÃO DE AUTORIA

Concepção: L. Silveira, M. K. Ferreira, S. E. Caregnato

Elaboração do manuscrito: L. Silveira, A. Barbosa, M. K. Ferreira

Coleta de dados: A. Barbosa, L. Silveira

Análise de dados: L. Silveira, A. Barbosa, M. K. Ferreira

Discussão dos resultados: L. Silveira, A. Barbosa, M. K. Ferreira

Revisão e aprovação: S. E. Caregnato

CONJUNTO DE DADOS DE PESQUISA

Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no artigo e na seção "Materiais suplementares".

FINANCIAMENTO

O artigo se insere nas atividades do projeto "Citação a dados de pesquisa: implicações práticas e teóricas para a comunicação científica", que recebeu Auxílio a Pesquisa do CNPq (Processo 431034/2018-4). Além disso, foi utilizada a estrutura da UFRGS e da Faculdade de Biblioteconomia e Comunicação, em especial o Centro de Documentação de Acervo Digital da Pesquisa (CEDAP).

CONSENTIMENTO DE USO DE IMAGEM

Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA

Não se aplica.

CONFLITO DE INTERESSES

Não se aplica.

LICENÇA DE USO

Os autores cedem à **Encontros Bibli** os direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a [Licença Creative Commons Attribution](#) (CC BY) 4.0 International. Esta licença permite que **terceiros** remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico. Os **autores** têm autorização para assumir contratos adicionais separadamente, para distribuição não exclusiva da versão do trabalho publicada neste periódico (ex.: publicar em repositório institucional, em site pessoal, publicar uma tradução, ou como capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

PUBLISHER

Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no [Portal de Periódicos UFSC](#). As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES

Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO

Recebido em: 12/03/2020 – Aprovado em: 11/05/2020 – Publicado em: 10/07/2020