

10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



CITAÇÃO A CONJUNTOS DE DADOS DE PESQUISA EM BIODIVERSIDADE BRASILEIRA: mapeamento a partir do GBIF e *OpenAlex*

Júlia Aquino Rosa

<https://orcid.org/0009-0000-8218-2992>
julia.aquino@ufrgs.br
Universidade Federal do Rio Grande do
Sul (UFRGS)

Pedro Henrique da Silva Rodrigues

<https://orcid.org/0000-0001-5909-2067>
pedrohenrique@ufrgs.br
Universidade Federal do Rio Grande do
Sul (UFRGS)

Lucas George Wendt

<https://orcid.org/0000-0002-4901-6826>
lucas.george.wendt@gmail.com
Universidade Federal do Rio Grande do
Sul (UFRGS)

Dra. Sônia Elisa Caregnato

<https://orcid.org/0000-0002-5676-2763>
sonia.caregnato@ufrgs.br
Universidade Federal do Rio Grande do
Sul (UFRGS)

Resumo:

Este trabalho investiga a reutilização de dados de pesquisa das Ciências Biológicas publicados no Brasil através das citações, utilizando conjuntos publicados no repositório Global Biodiversity Information Facility (GBIF) e com verificação de citações na base de dados OpenAlex. Os resultados indicam que, embora a infraestrutura tecnológica para o compartilhamento de dados esteja disponível, apenas uma pequena parcela é formalmente citada. Conclui-se que o reúso de dados apresenta crescimento moderado, evidenciando o potencial para acelerar a produção de conhecimento e fortalecer a colaboração científica.

Palavras-Chave: Dados de pesquisa. Citação. OpenAlex. GBIF.

EIXO TEMÁTICO:

Acesso Aberto, Ciência Aberta e Dados Abertos

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.1033

COMO CITAR ESTE ARTIGO:

ROSA, Júlia Aquino; WENDT, Lucas George; RODRIGUES, Pedro Henrique da Silva; CAREGNATO, Sônia Elisa. Citação a conjuntos de dados de pesquisa em biodiversidade brasileira: mapeamento a partir do GBIF e OpenAlex. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1033. Disponível em: <https://doi.org/10.22477/x.ebbc.1033>.



1 INTRODUÇÃO

Com a consolidação do movimento da Ciência Aberta, a temática acerca do compartilhamento de dados de pesquisa passou a integrar a agenda de pesquisadores de diferentes áreas do conhecimento. O objetivo de tais práticas é ampliar o acesso às pesquisas ao longo de todo o seu ciclo de vida, promovendo maior confiabilidade das informações científicas por meio do fortalecimento da reprodutibilidade, da replicabilidade e do reuso dos dados. O reuso, segundo Sayão e Sales (2014), amplia o potencial cognitivo dos dados, já que o valor da pesquisa deixa de estar restrito aos seus resultados originais e passa a se vincular à possibilidade de reinterpretar os dados em novos contextos, disciplinas ou abordagens, estimulando a colaboração para além de fronteiras geográficas e disciplinares.

Observa-se que a distinção entre uso e reuso de dados de pesquisa parte da ideia de que o uso corresponde ao emprego dos dados no âmbito da pesquisa para o qual eles foram originalmente coletados, tipicamente pelos próprios pesquisadores que os elaboraram. Já o reuso seria o reaproveitamento dos dados em novas pesquisas e/ou para propósitos distintos, tanto pelos próprios autores quanto por terceiros (Caregnato *et al.*, 2021). Embora o compartilhamento de dados seja considerado benéfico, ainda não é amplamente realizado por inúmeras razões, entre elas a falta de reconhecimento pelo enorme esforço de curadoria necessário para disponibilizá-los (Wilkinson *et al.*, 2016; Wood-Charlson *et al.*, 2022). Uma das formas de promover o compartilhamento de dados e, ao mesmo tempo, dar crédito aos autores é verificando o reuso dos dados de pesquisa (ou *datasets*), embora essa não seja uma atividade trivial. Ainda que existam indicadores de acesso e *download* úteis para verificar se os dados de pesquisa estão sendo acessados e potencialmente lidos por outros pesquisadores, sozinhos eles não mostram se, de fato, houve a reutilização dos dados em outras pesquisas.

Por outro lado, os produtos científicos tradicionais (artigos, relatórios de pesquisa, etc.), dispõem das citações não apenas como um indicador de acesso e leitura da obra, mas também como algo que expõe a incorporação e reconhecimento daquele trabalho, mostrando-se importante e relevante a ponto de ser citado por outros pesquisadores. Da mesma forma, supõe-se que o uso e o reuso de dados de pesquisa também podem ser avaliados através da citação formal dos dados em outros produtos científicos, principalmente nos artigos. Dessa perspectiva, sinalizar que esses dados fizeram parte do processo de pesquisa científica opera como um indicador de circulação e reuso, contribuindo para indicar relevância e impacto no ecossistema da Ciência Aberta. Além disso, pode-se considerar que as citações aos dados de pesquisa em artigos derivados refletem seu uso, enquanto sua citação em artigos de outros autores reflete o reuso.

Independentemente do debate sobre se a citação implica em reconhecer, respeitar, valorizar, dialogar criticamente ou refutar produções anteriores, há relativo consenso de que citar sinaliza um tipo de uso no âmbito da comunidade científica. Assim, com o objetivo de investigar se o uso ou reutilização de dados de pesquisa se manifesta por meio da citação, este trabalho busca identificar a incidência de citações aos conjuntos de dados científicos da área das Ciências Biológicas publicados no Brasil. As Ciências Biológicas foram escolhidas por representarem uma área em que há tradição na disponibilização de dados de pesquisa e, portanto, em que se poderia recorrer a um repositório consolidado, com possibilidade de apresentar registros de reuso por meio da citação.

A literatura que busca analisar padrões de citação de dados biológicos indica a existência de poucas normas consolidadas, resultando em práticas heterogêneas que variam, inclusive, para um mesmo tipo de dado. Bousfield *et al.* (2016) observaram que, por exemplo, alguns conjuntos de dados tendem a ser citados

formalmente nas listas de referências, enquanto outros recursos são mais frequentemente mencionados por meio de identificadores inseridos diretamente no corpo do texto dos artigos. Estudos baseados em mineração de texto na literatura biológica mostram de forma consistente que menções informais às bases de dados e identificadores no corpo dos artigos superam amplamente as citações estruturadas; além disso, em diferentes domínios científicos, menos de 10% dos artigos que utilizam dados incluem referências formais aos conjuntos de dados em suas listas bibliográficas (Byers *et al.*, 2024; Park *et al.*, 2018; Mayo *et al.*, 2016; Bousfield *et al.*, 2016; Kafkas *et al.*, 2013).

Quando adotados, os padrões técnicos de citação baseiam-se em identificadores persistentes, especialmente em Identificadores de Objeto Digital (DOIs), que permitem a vinculação entre publicações e conjuntos de dados e sustentam os princípios FAIR (Wilkinson *et al.*, 2016; Wood-Charlson *et al.*, 2022). Apesar da infraestrutura necessária já estar disponível, o principal obstáculo parece estar na limitada incorporação dessas práticas pela comunidade científica, mais do que na ausência de ferramentas (Corpas *et al.*, 2018; McMurry *et al.*, 2017).

2 METODOLOGIA

Para identificar a incidência de citações a conjuntos de dados de pesquisa das Ciências Biológicas publicados no Brasil, foram utilizadas duas fontes: a) o repositório da *Global Biodiversity Information Facility* (GBIF), cujo objetivo é dar acesso aos registros sobre espécies observadas e/ou coletadas; b) a OpenAlex, que indexa uma parte significativa da produção científica global, inclusive dados de pesquisa, recuperando suas referências e citações. As fontes permitiram, respectivamente, recuperar os metadados de conjuntos de dados de pesquisa brasileiros publicados em acesso aberto e contabilizar as citações recebidas por esses conjuntos.

A busca no GBIF foi realizada na coleção “*Datasets*”, utilizando o campo “*Publishing country or area*”, onde foi inserido o termo “*Brazil*”. A coleta foi realizada em dezembro de 2025 e resultou em 811 registros, que foram baixados em formato TSV. Na planilha resultante havia 17 colunas para cada registro, mas para este trabalho, foi utilizada somente a coluna “DOI”, pois não havia data de publicação ou métricas de citação. De posse dos DOIs, realizou-se uma consulta via API à OpenAlex que recuperou baixo volume de citações para poucos conjuntos e registros zerados e incompletos em sua maioria. Em vista disso, procedeu-se à busca manual através de cada DOI na OpenAlex em janeiro de 2026. Foram recuperados os seguintes campos de metadados: data de publicação, número de citações, número de documentos citantes (“*cited by*”) e tipo de publicação. Após a checagem, limpeza e padronização dos registros, o corpus final foi composto por 791 conjuntos de dados com DOI, publicados entre 2007 e 2025. O intervalo 2007-2025 refere-se exclusivamente ao ano de publicação dos conjuntos no GBIF/OpenAlex. As citações foram contabilizadas na OpenAlex até janeiro de 2026 (data da coleta).

3 RESULTADOS E DISCUSSÃO

O total de citações recuperadas para os 791 conjuntos foi de 239, sendo que apenas 75 (9,48%) receberam pelo menos uma citação. Desses, os dois mais citados receberam, coincidentemente, 57 citações cada: Flora e Funga do Brasil - Lista Oficial, conjunto cujo DOI é 10.15468/1mtkaw e Catálogo Taxonômico da Fauna do Brasil, DOI 10.15468/c4cauy. Tratam-se de grandes conjuntos de dados, produzidos em colaboração entre muitos pesquisadores. Após esses, três conjuntos receberam 8 citações cada, um recebeu 6, outro 4, sete receberam 3, nove receberam 2 e 52 receberam uma citação cada. Os restantes 716 conjuntos não receberam citações.

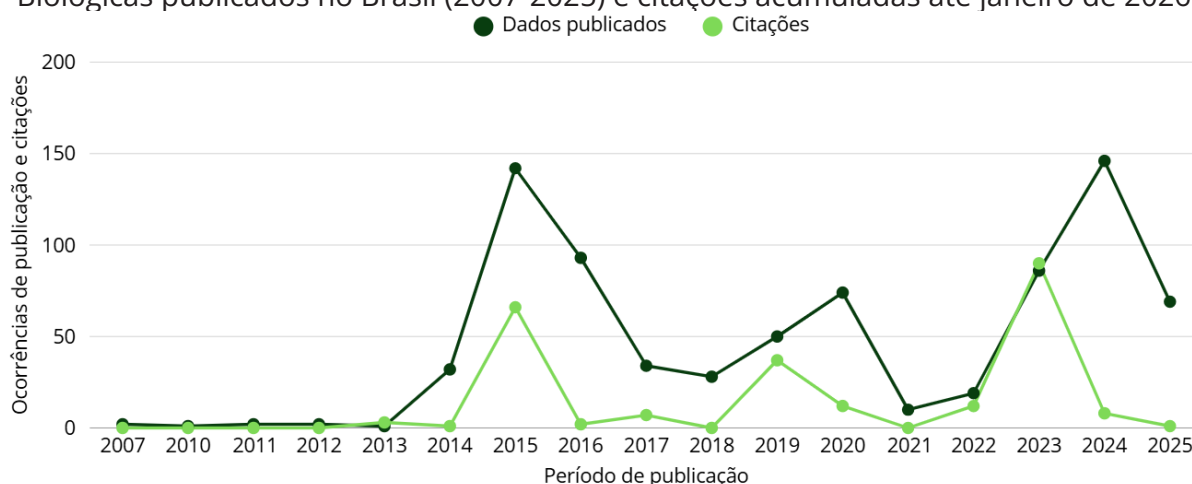
Observou-se uma expansão progressiva na publicação de dados ao longo do período, com concentração nos anos mais recentes, especialmente em 2024 que corresponde a 18,4% do total (Gráfico 1). Este comportamento pode insinuar um crescimento relativamente tardio da adesão de instituições brasileiras às práticas de publicação de dados com DOI, possivelmente associado à consolidação, por meio da adoção, de políticas de Ciência Aberta, além da ampliação de infraestruturas de repositórios e a valorização institucional da gestão e compartilhamento de dados de pesquisa, assim como o crescimento de solicitações das agências financiadoras e editores.

A despeito do crescimento quantitativo, tendo as citações tomadas como indicadores de uso e reúso, percebe-se que o reconhecimento dos conjuntos de dados permanecem baixos. Esse resultado é convergente com as evidências apontadas na literatura acerca da limitada incorporação das citações formais aos dados nas práticas correntes da comunicação científica. O fato da maioria absoluta dos conjuntos não apresentar qualquer citação sugere que a publicação de dados, embora crescente, ainda não se traduz de forma consistente em reutilização explícita ou reconhecimento formal por meio de citações, o que pode estar relacionado tanto à baixa cultura de citar os dados quanto a problemas de visibilidade, padronização ou mesmo rastreabilidade.

Embora a maioria dos conjuntos não tenha recebido citações, observa-se que os dois itens mais citados correspondem a produtos de referência nacional na área de biodiversidade e taxonomia (flora/funga e fauna). Esses resultados sugerem que a visibilidade e o reúso podem estar associados a três fatores combinados: maior quantidade de colaboração científica, relevância temática e tipologia documental de caráter estruturante, frequentemente utilizada como fonte de referência por diferentes estudos. Assim, a concentração de citações não deve ser

interpretada apenas como efeito do volume de dados publicados, mas também das características e funções do tipo de dado disponibilizado. Nesta pesquisa, tais fatores foram examinados de forma descritiva, portanto, não é possível afirmar relação causal entre coautoria, temática e impacto de citação.

Gráfico 1: Frequência anual de publicação de conjuntos de dados na área das Ciências Biológicas publicados no Brasil (2007-2025) e citações acumuladas até janeiro de 2026.



Fonte: Elaborado pelos autores (2026).

A análise temporal das citações, Gráfico 1, indica um possível descompasso entre o pico de publicação de dados e o momento em que passam a ser efetivamente reutilizados e citados. O intervalo temporal é consistente com o ciclo de produção científica, no qual a coleta, análise, redação e publicação de artigos e outros documentos que reutilizam dados tendem a ocorrer após a disponibilização dos conjuntos. Além disso, a predominância de artigos científicos como principal tipo de documento citante (9,10%) indica que a circulação dos dados ocorre majoritariamente no circuito formal da literatura acadêmica, com baixa penetração em outros tipos de produtos científicos ou técnicos.

As recomendações atuais indicam que as citações adequadas aos dados de pesquisa devem incluir menções claras no texto e referência formal completa na lista de referências, além de serem citados em campo exclusivo nos artigos que descrevem originalmente os conjuntos de dados. Recomenda-se ainda que os DOIs

sejam apresentados como URIs, garantindo acesso persistente e reprodutibilidade da pesquisa (Penev *et al.*, 2017).

Entre os problemas encontrados no uso das bases de dados, estão inconsistências entre as classificações do GBIF e da OpenAlex no que se refere ao campo de metadados “tipos de publicação”: o GBIF, que deveria oferecer apenas conjuntos de dados, recuperou 1,6% de DOIs vinculados a artigos científicos, que, por isso, foram excluídos da pesquisa. Em contrapartida, a OpenAlex apresentou uma fração menor desta inconsistência, classificando como artigo e *review* apenas 0,25% e 0,12%, embora a verificação manual tenha confirmado que se tratava efetivamente de dados de pesquisa. As discrepâncias impactam a contagem e a caracterização dos conjuntos analisados, bem como também revelam problemas que perpassam os sistemas automáticos de tipificação documental, com implicações diretas para estudos métricos e para a própria avaliação do uso e impacto de dados de pesquisa.

4 CONSIDERAÇÕES FINAIS

É possível verificar o crescimento da disponibilização de conjuntos de dados e de citações, mesmo que de forma moderada. No entanto, com base nos resultados destacados acima, ainda não é possível argumentar que a citação aos dados de pesquisa seja um indicador de uso, semelhantemente ao que acontece com a literatura científica, mesmo em áreas que têm tradição de disponibilizar em repositórios abertos seus conjuntos de dados, como as Ciências Biológicas.

Os incentivos para o compartilhamento e disponibilização aberta de dados de pesquisa ainda precisam ser realizados de forma incisiva, assim como é necessário destacar o potencial de reúso dos dados na produção de novas pesquisas científicas, tornando o processo mais rápido e colaborativo.

Quando os dados são reconhecidos como fontes primárias de conhecimento, passam a ocupar um lugar de destaque na comunicação científica, permitindo que sejam analisados, reinterpretados e debatidos sob múltiplas perspectivas e com objetivos de pesquisa distintos. Nesse sentido, o reuso de dados fortalece a ciência e amplia o impacto das pesquisas, permitindo o reconhecimento e colaboração.

Na continuidade deste trabalho, buscar-se-á identificar os tipos de fontes citantes, além da disposição espacial das citações no texto científico.

REFERÊNCIAS

CAREGNATO, S. E.; ROCHA, R. P. da; GABRIEL JÚNIOR, R. F. Reuso de dados: princípios FAIR e o ecossistema de pesquisa. In: SALES, L. F. et al. (org.) **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 187–200. DOI: 10.22477/9786589167242.cap14. Acesso em: 17 abril de 2026

CORPAS, M.; KOVALEVSKAYA, N. V.; MCMURRAY, A.; NIELSEN, F. G. G. A FAIR guide for data providers to maximise sharing of human genomic data. **PLoS Computational Biology**, v. 14, n. 3, 2018. DOI: <https://doi.org/10.1371/journal.pcbi.1005873>. Acesso em: 17 abril de 2026

BOUSFIELD, D.; MCENTYRE, J.; VELANKAR, S.; PAPADATOS, G.; BATEMAN, A.; COCHRANE, G.; KIM, J. H.; GRAEF, F.; VARTAK, V.; ALAKO, B.; BLOMBERG, N. Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. **F1000Research**, v. 5, 11 fev. 2016. DOI: 10.12688/f1000research.7911.1. Acesso em: 17 abril de 2026

BYERS, N.; PARKER, C.; BEECROFT, C.; REDDY, T. B. K.; SALAMON, H.; GARRITY, G.; FAGNAN, K. Identifying genomic data use with the Data Citation Explorer. **Scientific Data**, v. 11, n. 1200, 2024. DOI: <https://doi.org/10.1038/s41597-024-04049-7>. Acesso em: 17 abril de 2026

KAFKAS, S.; KIM, J. H.; MCENTYRE, J. R. Database Citation in Full Text Biomedical Articles. **PLoS ONE**, v. 8, n. 5, 2013. DOI: <https://doi.org/10.1371/journal.pone.0063184>. Acesso em: 17 abril de 2026

MAYO, C.; VISION, T. J.; HULL, E. A. The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. **International Journal of Digital Curation**, v. 11, n. 1, p. 150-155, 2016. DOI: <https://doi.org/10.2218/ijdc.v11i1.400>. Acesso em: 17 abril de 2026

MCMURRY, J. A.; JUTY, N.; BLOMBERG, N.; BURDETT, T.; CONLIN, T. *et al.* Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. **PLoS Biology**, v. 15, n. 6, 2017. DOI: <https://doi.org/10.1371/journal.pbio.2001414>. Acesso em: 17 abril de 2026

PARK, H.; YOU, S.; WOLFRAM, D. 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, 2018. DOI: 10.1002/ASI.24049 Acesso em: 17 abril de 2026

PENEV, L.; MIETCHEN, D.; CHAVAN, V. S.; HAGEDORN, G.; SMITH, V. S. *et al.* Strategies and guidelines for scholarly publishing of biodiversity data. **Research Ideas and Outcomes**, v. 3, 2017. DOI: <https://doi.org/10.3897/rio.3.e12431>. Acesso em: 17 abril de 2026

SAYÃO, L. F.; SALES, L. F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, v. 8, n. 2, p. 76–92, jun. 2014. Disponível em: <https://doi://10.3395/reciis.v8i2.934.pt>. Acesso em: 17 abril de 2026.

WILKINSON, M.D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G. AXTON, M. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v.3, 2016. Disponível em: <https://doi.org/10.1038/sdata.2016.18>. Acesso em: 17 abril de 2026.

WOOD-CHARLSON, E. M.; CROCKETT, Z.; ERDMANN, C.; ARKIN, A. P.; ROBINSON, C. B. Ten simple rules for getting and giving credit for data. **PLoS Computational Biology**, v. 18, n. 9, 2022. Disponível em: <https://doi.org/10.1371/journal.pcbi.1010476>. Acesso em: 17 abril de 2026.