Gestão de dados científicos: produção e impacto a partir de dados da base *Dimensions*

Research data management: production and impact from Dimensions database data

Marília Catarina Andrade Gontijo
Universidade Federal de Minas Gerais, Brasil.
E-mail: mariliacgontijo@gmail.com
ORCID: https://orcid.org/0000-0002-9181-0302.

Raíssa Yuri Hamanaka
Universidade Estadual de Londrina, Brasil.
E-mail: raissa.hamanaka@uel.br
ORCID: https://orcid.org/0000-0001-9516-5825

Ronaldo Ferreira de Araújo
Universidade Federal de Alagoas, Brasil.
E-mail: ronaldo.araujo@ichca.ufal.br
ORCID: http://orcid.org/0000-0003-0778-9561

RESUMO

A pesquisa objetivou analisar a produção científica sobre gestão de dados científicos indexada na *Dimensions*. A partir da busca pelo termo *"research data management"* foram recuperados 677 artigos, analisados por meio de indicadores bibliométricos de produção e citação. A multidisciplinaridade em gestão de dados de pesquisa foi demonstrada pelas publicações ocorrerem em diferentes áreas de pesquisa, como *Information and Computing Sciences, Information Systems, Library and Information Studies, Medical and Health Sciences e History and Archaeology.* Os países com maiores índices de publicações foram Estados Unidos, Alemanha e Reino Unido. Cerca de 60% das publicações tiveram pelo menos uma citação, com um total de 3.598 citações encontradas, caracterizando-se um impacto acadêmico crescente uma vez que o volume de produção e de citações têm crescido ao longo do tempo. Ao pensar

Como citar: Gontijo, M. C. A.; & Hamanaka, R. Y.; & Araújo, R. F. (2022). Gestão de dados científicos: produção e impacto a partir de dados da base *Dimensions.* En T. M. R. Dias (Ed.), Informação, Dados e Tecnologia. *Advanced Notes in Information Science, volume 2* (pp. 112-120). Tallinn, Estonia: ColNes Publishing. DOI: 10.47909/anis.978-9916-9760-3-6.89.

na era do *Big Data*, a gestão de dados é um tema em desenvolvimento para garantir o compartilhamento e reuso destes, e consequentemente, o avanço da ciência. Desta forma, este estudo bibliométrico permitiu acompanhar o desempenho da literatura sobre gestão de dados científicos.
Palavras-chave: gestão de dados científicos, bibliometria, produção científica, ciência aberta.

ABSTRACT

The study aims to analyze the scientific production of research data management indexed in Dimensions. Using the term "research data management", 677 articles were retrieved and analyzed using output and citation bibliometric indicators. The multidisciplinary in research data management was demonstrated by publications occurring in different research areas, such as computer science, information systems, library and information science, medicine and health sciences, and history and archeology. The countries with the highest publication rates were the United States, Germany, and the United Kingdom. About 60% of the publications had at least one citation, with 3,598 citations found, featuring a growing academic impact since the volume of production and citations have grown over time. When it comes to the Big Data era, data management is a topic under development that ensures its sharing and reuse and, consequently, the advancement of science. This bibliometric study made it possible to monitor the literature performance on research data management.
Keywords: *research data management, bibliometrics, scientific production, open science*

INTRODUÇÃO

A ciência aberta vem se desenvolvendo, e pode ser entendida como um macro termo que engloba o acesso aberto, dados científicos abertos, cadernos de pesquisa abertos, avaliação científica aberta, educação aberta, ferramentas e materiais científicos abertos, ciência cidadã e políticas de ciência aberta (Albagli, Clinio & Raychtock, 2014; Pontika *et al.*, 2015). Segundo Rodrigues *et al.* (2017) a gestão de dados pode ser definida como uma disciplina que permite armazenar, utilizar e analisar dados, por meio da implementação de políticas, estratégias, procedimentos e práticas que mantenham a integridade, segurança, padronização e organização dos dados.

A gestão de dados de pesquisa "[...] é o desenvolvimento e implementação de políticas, planos e processos que gerenciem esses dados para manter sua integridade, segurança e usabilidade" (Specht *et al.*, 2015, p. 145, tradução nossa). Di Martino *et al.* (2014) definem uma cadeia de valor necessária à gestão de grandes volumes de dados: aquisição dos

dados, análise dos dados, curadoria dos dados, armazenamento dos dados e uso dos dados. Essa cadeia pode ser entendida como um conjunto de fases necessárias à gestão de dados de pesquisa. Seu objetivo é permitir que os dados científicos sejam autodescritivos, para que possam ser efetivamente reutilizados quando forem descobertos (Specht *et al.,* 2015). Hey e Trefethen (2013) caracterizam o dilúvio de dados e como a infraestrutura científica precisará se adequar aos mecanismos de produção, compartilhamento, curadoria, preservação e reuso dos dados científicos. Os autores caracterizam o processo de automação da gestão de dados, automatizando-se a descoberta científica, com dados gerados por meio de simulações, experimentos e sensores sendo manipulados, visualizados e analisados com apoio de intensa infraestrutura tecnológica.

No paradigma da ciência orientada a dados, vem se tornando comum a exigência do depósito de conjuntos de dados coletados antes mesmo da publicação dos artigos, assim como a exigência por algumas agências de fomento do Plano de Gestão de Dados para que um projeto seja aprovado. Com esse aumento do interesse, e consequentemente, do volume de investigações sobre gestão de dados científicos, esta pesquisa objetiva averiguar o desempenho da produção científica sobre essa temática. Para tanto, são aplicados estudos de produtividade científica, como os métricos da informação, que tem como um de seus subcampos a bibliometria. Estudo métrico que pode ser caracterizado pelo uso de métodos estatísticos e matemáticos na quantificação, descrição e diagnóstico de comunicações escritas (Guedes & Borschiver, 2005), considerado um instrumento de análise sistemática de publicações (Kalantari *et al.,* 2017).

OBJETIVO

A presente pesquisa tem como objetivo analisar a produção e o impacto da produção científica internacional sobre gestão de dados científicos indexada na base de dados *Dimensions* (https://www.dimensions.ai) por meio do uso da bibliometria.

Por meio dos indicadores bibliométricos, é possível mapear o perfil da produção científica dos diversos campos do conhecimento, averiguando suas características e seu impacto acadêmico. Dessa forma, permite-se, também, que se atribua visibilidade aos seus componentes, como: autores, periódicos científicos, organizações e países, ao utilizar indicadores bibliométricos que acompanham o desempenho da produtividade das publicações científicas (Vanti & Sanz-Casado, 2016). O indicador de citação, por exemplo, é utilizado para analisar o impacto acadêmico de pesquisas (Freitas *et al.,* 2017), ao quantificar dados de citações, como autor, título, origem geográfica, ano e idioma da publicação (Foresti, 1990).

PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa de caráter descritivo e exploratório, com abordagem quantitativa, que objetiva analisar a produção científica internacional sobre gestão de dados científicos na base de dados *Dimensions* a partir de indicadores bibliométricos. Optou-se por essa base por apresentar uma ampla cobertura de dados internacionais, abrangendo variados tipos de publicações, dentre as quais artigos científicos, em diferentes campos de pesquisa e regiões, disponibilizando um grande volume de dados para análises (Bode *et al.*, 2019).

Para a busca, utilizou-se a palavra-chave *"research data management"* entre aspas para a recuperação exata do termo, por representar a tradução no idioma inglês para gestão de dados científicos. Dentre os estudos bibliométricos correlatos estão: Guimarães e Bezerra (2019) que utilizaram o termo *"data management"* e Zhang e Eichmann-Kalwara (2019) com *"research data management"*.

A busca foi realizada nos campos de título e resumo, por, de maneira geral, apresentarem a ideia central dos estudos, servindo como orientação sobre o assunto aos leitores (Costa & Moura, 2013; Creswell, 2010). Foi utilizado o filtro por artigos, por serem uma das fontes de disseminação de pesquisas mais escolhidas pelos autores (Maricato & Lima, 2017). Não foi realizado recorte temporal, sendo considerados todos os anos de publicação disponibilizados pela *Dimensions* até agosto de 2021, período de realização da busca bibliográfica.

A partir da aplicação desses filtros foram recuperados 677 artigos, os quais foram analisados segundo seus indicadores bibliométricos de produtividade (ano de publicação, países/regiões, organizações, autores, periódicos científicos e campos de pesquisa mais produtivos) e citação (total de citação obtida pelo conjunto analisado e *ranking* de artigos mais citados).

RESULTADOS

De acordo com os dados da pesquisa houve variação no aumento das publicações ao longo dos anos (1970 a 2021). Entre 1970 e 2011 manteve-se uma variação de uma a sete publicações, enquanto em 2012 houve um salto em comparação aos anos anteriores, com 22 publicações. A partir de 2012 houve grande crescimento da literatura sobre gestão de dados científicos, tendo como ápice das produções os anos de 2019 (116 artigos, 17% do universo) e 2020 (102 artigos, 15%). O aumento nas publicações sobre o tema já havia sido evidenciado nos estudos bibliométricos de Guimarães e Bezerra (2019) e Zhang e Eichmann-Kalwara (2019).

Os países/regiões com maiores números de publicações na temática foram Estados Unidos da América (EUA), Alemanha e Reino Unido. Os EUA apresentaram 19,64% das publicações do universo, com 133 artigos, tendo como organizações mais produtivas *Oregon State University* (sete artigos), *Stanford University* (seis) e *University of Michigan* e *University of Washington* (cinco artigos cada). A Alemanha foi responsável por 13,44% das publicações, com 91 artigos, tendo como principais organizações produtoras: *University of Göttingen* (nove artigos), *University Medical Center Göttingen* e *University of Cologne* (seis artigos cada) e *RWTH Aachen University* (cinco artigos). E o Reino Unido apresentou 7,8% das publicações, com 53 artigos, com o monopólio destas concentrado nas organizações de pesquisa: *University of Sheffield* (12 artigos), *University of Oxford* (quatro) e *Imperial College London* (três).

Os autores que publicaram mais de cinco artigos foram Andrew Martin Cox da *University of Sheffield* no Reino Unido (13 artigos), Stephen Pinfield da *University of Sheffield* e Marta Teperek da *Delft University of Technology na Holanda* (sete publicações cada) e Mary Anne Kennan da *Charles Sturt University* na Austrália e Amanda L. Whitmire da *Stanford University* nos EUA (seis publicações cada). Enquanto os demais autores apresentaram entre um e cinco artigos. Esse resultado foi semelhante ao de Guimarães e Bezerra (2019), corroborando a lei de Lotka, ao estabelecer que poucos autores produzem muito e muitos autores produzem pouco (Costa & Cunha, 2015; Guedes & Borschiver, 2005).

Os periódicos científicos com mais publicações sobre o tema foram: *International Journal of Digital Curation* com 74 artigos, *Journal of eScience Librarianship* com 38, *Data Science Journal* com 24, *Septentrio Conference Series* com 15 e *Bulletin of the Association for Information Science and Technology* com 13. Os campos de pesquisa (denominados e agrupados pela base de dados *Dimensions)* com maiores números de publicações não exclusivas, podendo ser incluídas em mais de um deles, foram: *Information and Computing Sciences* com 482 artigos, *Information Systems* com 360, *Library and Information Studies* com 113, *Medical and Health Sciences* com 53, e *History and Archaeology* com 35. Esses resultados indicam a multidisciplinaridade presente na gestão de dados científicos, também apontada por Zhang e Eichmann-Kalwara (2019) e Cunha e Costa (2020). É possível indagar se a multidisciplinaridade está relacionada com o fato da coleta, do uso e do compartilhamento de dados serem processos realizados em diversos campos do conhecimento.

De acordo com as análises do indicador de citação, dos 677 artigos que compõem o universo da pesquisa, 415 (61%) apresentaram uma ou mais citações, havendo um total de 3.598 citações. Houve crescimento a partir de 2010 com 28 citações, 2015 com 189, 2020 com 605 e 2021 com 660

(até agosto desse ano). O aumento no número de citações ocorreu à medida que houve o crescimento na quantidade de publicações, como também apontado por Guimarães e Bezerra (2019) e Cunha e Costa (2020).

Entre os 10 artigos mais citados, apresentados na Tabela 1, têm-se:

AUTORES	TÍTULO DO ARTIGO	CITAÇÕES
Lowe *et al.* (2009)	STRIDE--An integrated standards-based translational research informatics platform	306
Sharma *et al.* (2014)	Panorama: A Targeted Proteomics Knowledge Base	144
Corrall *et al.* (2013)	Bibliometrics and Research data management Services: Emerging Trends in Library Support for Research	104
Tenopir *et al.* (2014)	Research data management services in academic research libraries and perceptions of librarians	95
Cox e Pinfield (2013)	Research data management and libraries: Current activities and future priorities	94
Anderson *et al.* (2007)	Issues in Biomedical Research data management and Analysis: Needs and Barriers	89
Skripcak *et al.* (2014)	Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymized public datasets	76
Cox *et al.* (2017)	Developments in research data management in academic libraries: Towards an understanding of research data service maturity	66
Arend *et al.* (2016)	PGP repository: a plant phenomics and genomics data publication infrastructure	61
Arend *et al.* (2014)	e!DAL - a framework to store, share and publish research data	57

Tabela 1. Artigos científicos mais citados. **Fonte:** dados da pesquisa, 2021).

Percebe-se que os três primeiros colocados apresentam quantidades de citações com elevada diferença em relação ao restante. As temáticas dos artigos mais citados variam de plataformas, *frameworks* e repositórios que permitam o armazenamento, compartilhamento e reuso de dados até barreiras e serviços relacionados a gestão de dados científicos.

CONSIDERAÇÕES FINAIS

Objetivou-se analisar a produção científica internacional sobre gestão de dados científicos indexada na *Dimensions* a partir de indicadores bibliométricos de produtividade e citação, utilizando como termo de busca *"research data management"*.

A partir da análise das pesquisas recuperadas, observou-se um crescimento da produção científica ao longo dos anos, com destaque a partir de 2012 e pico de publicações em 2019 e 2020 (116 e 102 artigos respectivamente). Os países/regiões com maiores índices de publicações foram os EUA (133 artigos), Alemanha (91) e Reino Unido (53). Entre os autores mais produtivos, se destacam Andrew Martin Cox (13 artigos), Stephen Pinfield e Marta Teperek (sete cada) e Mary Anne Kennan e Amanda L. Whitmire (seis cada). A multidisciplinaridade em gestão de dados de pesquisa foi demonstrada pela variedade de áreas de periódicos científicos e campos do conhecimento que publicam sobre a temática.

Ao pensar na era do *Big Data,* a gestão de dados é um tema em desenvolvimento para garantir o compartilhamento e reuso destes, e consequentemente, o avanço da ciência. Este estudo bibliométrico permitiu acompanhar o desempenho da literatura sobre gestão de dados científicos, ao delinear o perfil das produções científicas, o aumento das publicações ao longo dos anos, além da verificação da visibilidade gerada por pesquisadores, periódicos, organizações e regiões.

Uma limitação da pesquisa foi o estudo do desempenho exclusivamente acadêmico da produção científica na temática delimitada, não considerando o impacto social e a atenção *on-line* desta. Estes poderiam ser aferidos por meio de indicadores altmétricos. Como pesquisa futura sugere-se a ampliação das bases de dados utilizadas no estudo bibliométrico e a combinação dos indicadores bibliométricos e altmétricos. Para melhor qualificação temática da produção, será incluída em sua próxima etapa uma análise da coocorrência de termos presentes nos títulos e palavras-chave dos estudos.

CONFLITOS DE INTERESSE

Os autores declaram que não há conflitos de interesse.

DECLARAÇÃO DE CONTRIBUIÇÃO

Conceptualização, curadoria de dados, análise formal, pesquisa, metodologia, administração do projeto, validação, visualização, escrita do rascunho inicial e escrita da revisão e edição: Marília Catarina Andrade, Raíssa Yuri Hamanaka e Ronaldo Ferreira de Araújo.

DECLARAÇÃO DE CONSENTIMENTO DE DADOS

Os dados gerados durante este artigo foram incluídos no manuscrito.

AGRADECIMENTOS

A primeira autora agradece à Capes pelo apoio financeiro concedido no desenvolvimento de sua pesquisa de doutorado. Os autores agradecem à Dimensions pelo acesso e uso não comercial dos dados da produção científica e de citação utilizados no estudo.

REFERÊNCIAS

Albagli, S., Clinio, A., & Raychtock, S. (2014). Ciência Aberta: correntes interpretativas e tipos de ação. *RECIIS: Revista Eletrônica de Comunicação em Informação, Inovação e Saúde, 10*(2), 434-450. https://doi.org/10.18617/liinc.v10i2.749

Bode, C. *et al.* (2019). *A Guide to the Dimensions Data Approach*. Cambridge: Digital Science.

Costa, M. M., & Cunha, M. B. (2015). A literatura internacional sobre e-Science nas bases de dados LISA e LISTA. *Encontros Bibli: revista eletrônica de Biblioteconomia e Ciência da Informação, 20*(44), 127-144. https://doi.org/10.5007/1518-2924.2015v20n44p127

Costa, M. U. P. da; & Moura, M. A. (2013). A representação da informação em contextos de comunicação científica: a elaboração de resumos e palavras-chave pelo pesquisador-autor. *Informação & Informação, 18*(3), 45-67. http://dx.doi.org/10.5433/1981-8920.2013v18n3p45

Creswell, J. W. (2010). *Projeto de pesquisa: métodos qualitativo, quantitativo e misto* (3ª ed.). Porto Alegre: Artmed: Bookman.

Cunha, M. B., & Costa, M. M. (2020). Fontes de informação sobre gestão de dados de pesquisa. *Informação & Sociedade: Estudos, 30*(4), 1-59. https://doi.org/10.22478/ufpb.1809-4783.2020v30n4.57183

Di Martino, B., Aversa, R., Cretella, G., Esposito, A., & Kołodziej, J. (2014). Big data (lost) in the cloud. *International Journal of Big Data Intelligence,* 1(1-2), 3-17.

Foresti, N. A. B. (1990). Contribuição das revistas brasileiras de biblioteconomia e ciência da informação enquanto fonte de referência para a pesquisa. *Ciência da Informação, 19*(1), 53-71. Recuperado de http://revista.ibict.br/ciinf/article/view/375/375

Guedes, V. L. S., & Borschiver, S. (2005). *Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de*

informação, de comunicação e de avaliação científica e tecnológica. Artigo apresentado no 6º Encontro Nacional de Ciência da Informação, Salvador, Brasil. Recuperado de http://www.cinform-anteriores.ufba.br/vi_anais/docs/VaniaLSGuedes.pdf

Guimarães, A. J. R., & Bezerra, C. A. (2019). Gestão de dados: uma abordagem bibliométrica. *Perspectivas em Ciência da Informação, 24*(4), 171-186. https://doi.org/10.1590/1981-5344/4192

Hey, T., & Trefethen, A. (2003). The data deluge: an e-science perspective. In Berman, F., Fox, G., & Hey, A. J. G. (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 809-824). [S. l.]: Wiley.

Kalantari, A., Kamsin, A., Kamaruddin, H. S., Ebrahim, N. A., Gani, A., Ebrahimi, A., & Shamshirband, S. (2017). A bibliometric approach to tracking big data research trends. *Journal of Big Data, 4*(1), 1-18. https://doi.org/10.1186/s40537-017-0088-1

Freitas, J. L., Rosas, F. S., & Miguel, S. E. (2017). Estudos métricos da informação em periódicos do portal SciELO: visibilidade e impacto na Scopus e Web of Science. *Palabra Clave, 6*(2), 1-12. https://doi.org/10.24215/PCe021

Maricato, J. de M., & Lima, E. L. M. (2017). Impactos da Altmetria: aspectos observados com análises de perfis no Facebook e Twitter. *Informação & Sociedade: Estudos, 27*(1), 137-145. Recuperado de https://periodicos.ufpb.br/ojs/index.php/ies/article/view/30921/17418

Pontika, N., Knoth, P., Cancellieri, M., & Pearce, S. (2015, October). Fostering open science to research using a taxonomy and an eLearning portal. In *Proceedings of the 15th international conference on knowledge technologies and data-driven business* (pp. 1-8). https://doi.org/10.1145/2809563.2809571

Rodrigues, A. A., Nóbrega, E., & Dias, G. A. (2017). Desafios da gestão de dados na era do Big Data: perspectivas profissionais. *Informação & Tecnologia, 4*(2), 63-79. https://doi.org/10.22478/ufpb.2358-3908.2017v4n2.40538

Specht, A. *et al.* (2015). Data management challenges in analysis and synthesis in the ecosystem sciences. *Science of the Total Environment, 534,* 144-158. https://doi.org/10.1016/j.scitotenv.2015.03.092

Vanti, N., & Sanz-Casado, E. (2016). Altmetria: a métrica social a serviço de uma ciência mais democrática. *Transinformação, 28*(3), 349-358. https://doi.org/10.1590/2318-08892016000300009

Zhang, L., & Eichmann-Kalwara, N. (2019). Mapping the Scholarly Literature Found in Scopus on Research data management: A Bibliometric and Data Visualization Approach. *Journal of Librarianship and Scholarly Communication, 7*(1). https://doi.org/10.7710/2162-3309.2266