6.07.99 - Ciência da Informação

GESTÃO DE DADOS DE PESQUISA: UM ESTUDO BIBLIOMÉTRICO DA PRODUÇÃO CIENTÍFICA INDEXADA NA *WEB OF SCIENCE*

Marília C. A. Gontijo[1], Raíssa Y. Hamanaka[2]
1. Mestre em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação da Universidade Federal de Minas Gerais (ECI-UFMG)
2. Doutoranda em Ciência da Informação pelo Centro de Educação, Comunicação e Artes da Universidade Estadual de Londrina (CECA-UEL)

Resumo

A gestão de dados de pesquisa tem como objetivo organizar dados não-estruturados de forma a garantir sua preservação e recuperação ao longo do tempo, além de torná-los interpretáveis e permitir seu reuso. O objetivo da pesquisa é caracterizar e analisar o desempenho da produção científica sobre gestão de dados de pesquisa por meio de técnicas bibliométricas. Os metadados das publicações foram obtidos em consulta na base de dados *Web of Science,* entre 2017 e 2021, e elaborou-se um mapa bibliométrico por meio do *software VOSviewer*. Foram analisadas 486 publicações sobre gestão de dados de pesquisa. Os indicadores bibliométricos permitiram identificar que houve crescimento da produção científica no período pesquisado; os países com maiores números de publicações foram os Estados Unidos da América, seguido por Inglaterra e Alemanha; os autores que mais publicaram foram Ricardo-Adan Salas-Rueda, Rodrigo-David Salas-Rueda e Joern Loetsch; e os periódicos científicos com mais artigos publicados foram *IEEE Access*, *PLOS ONE* e *Sensors*. Este estudo bibliométrico permitiu entender o desempenho da literatura sobre gestão de dados de pesquisa, assim como compreender os padrões de pesquisa e as redes de colaboração entre os diversos atores sobre a temática.

Palavras-chave: estudos métricos da informação; bibliometria; produção científica; gestão de dados.

Introdução

A contemporaneidade está sendo marcada pelo grande volume de dados não-estruturados, que necessitam ser armazenados e acessados a rápida velocidade, fenômeno denomidado *Big Data* (BOERES; CUNHA, 2016). Concomitantemente, observa-se o surgimento de um novo paradigma na ciência denominado *e-Science*, orientado pela necessidade de armazenamento, compartilhamento e reuso dos dados para posterior análise e possível formulação de teorias (SAYÃO; SALES, 2012).

O estudo da gestão de dados de pesquisa contribui para as práticas de armazenamento, representação e descrição dos dados de pesquisa, permitindo que estes possam ser reutilizados em outros contextos, e consequentemente, avançando a ciência. Do ponto de vista econômico, a devida documentação de experimentos é de extrema importância, tendo em vista que alguns experimentos são simples e baratos, enquanto alguns são dispendiosos e complexos de replicar (SAYÃO; SALES, 2012). De maneira semelhante a bibliotecas, museus e arquivos, atualmente, os repositórios digitais funcionam como instituições de patrimônio intelectual, ao garantirem a preservação da memória institucional quando armazenam informações para uso corrente e futuro (SAYÃO; SALES, 2013). Além disso, a gestão dos dados de pesquisa permite o "[…] acesso livre aos produtos de pesquisas financiados com recursos públicos" (SAYÃO; SALES, 2016, p. 111).

Com o crescimento de pesquisas sobre o tema, surge a necessidade de estudos voltados a produtividade dessa área, que permitam compreender como se dá o desempenho de sua literatura, por meio dos estudos métricos da informação, como a bibliometria, que utiliza métodos matemáticos e estatísticos na descrição da produção científica, e dentre os parâmetros observáveis que utiliza estão: autores, citações, palavras-chave, periódicos, publicações e usuários (GUEDES, BORSCHIVER, 2005). Os estudos bibliométricos são utilizados para o levantamento das tendências, potencialidades e padrões de pesquisa dos variados campos do conhecimento, contribuindo para a melhoria da pesquisa científica (SACARDO; HAYASHI, 2013).

A partir da problematização exposta, a presente pesquisa tem como objetivo caracterizar e analisar o desempenho da produção científica sobre gestão de dados de pesquisa por meio de técnicas bibliométricas.

Metodologia

Esta pesquisa se caracteriza como descritiva, exploratória e bibliográfica. Possui abordagem quantitativa ao utilizar técnicas bibliométricas para levantamento e análise dos resultados. Foi realizada uma busca bibliográfica na *Web of Science* em fevereiro de 2021. Foram utilizados os termos "*data science*", "*data management*" e "*digital curation*", separados pelo operador booleano "OR" e aplicados no campo de título. O recorte temporal foi de 2017 a 2021, e os filtros utilizados foram artigos de periódicos e em acesso aberto. Foram recuperados metadados de 486 artigos, classificados por número de citações, e extraídos de acordo com os registros completos e referências citadas, para serem tabulados e inseridos no *software VOSviewer* (https://www.vosviewer.com/) para a geração de um mapa bibliométrico que permitisse a visualização das redes bibliométricas.

A produção científica sobre gestão de dados foi analisada por meio dos indicadores bibliométricos: ano de publicação, organizações, países e regiões, autores, periódicos científicos e áreas de pesquisa. Também foram criadas redes de coocorrências de palavras-chave. Além disso, foram identificados indicadores relacionados às citações: o total de citações do universo investigado, os artigos mais citados e a distribuição das citações por ano.

Foram encontrados estudos correlatos que utilizaram o *VOSviewer* (CODATO, 2018; LIMA; LEOCÁDIO, 2018; PALLUDETO; FELIPINI, 2019; SOUSA *et al*., 2019) e, estudos bibliométricos sobre gestão de dados (COSTA; CUNHA, 2015; GUIMARÃES; BEZERRA, 2019; ZHANG; EICHMANN-KALWARA, 2019), que fundamentaram as discussões dos resultados levantados nesta pesquisa.

Resultados e Discussão

As 486 publicações foram analisadas sob o enfoque dos indicadores bibliométricos. Em relação à distribuição das publicações por ano, notou-se o crescimento da quantidade de publicações no período analisado (2017-2021), com salto nas publicações em 2019 (32,3%), apresentando menor quantidade de artigos em 2021, sete, por ser um ano ainda em seu início. O aumento das publicações em torno da gestão de dados já havia sido evidenciado nos estudos bibliométricos de Costa e Cunha (2015), Guimarães e Bezerra (2019) e Zhang e Eichmann-Kalwara (2019). O crescimento das publicações sobre a temática também foi evidenciado nas bibliografias de Cunha e Costa (2020) e Szigeti e Wheeler (2011).

Os países que tiveram mais publicações foram Estados Unidos da América (EUA), seguido por Inglaterra e Alemanha. O estudo bibliométrico de Costa e Cunha (2015) em bases de dados da Ciência da Informação apontou o mesmo resultado. Na presente pesquisa, os EUA tiveram aproximadamente 30% das publicações da amostra, com 160 artigos, tendo como organizações mais produtivas a *University of California*, com 25 artigos, seguida pelo *System United States Department of Energy*, com 17, e pela *Harvard University,* com 13. As organizações inglesas com mais publicações foram a *University of Manchester*, com oito artigos e a *University of Oxford*, com sete, tendo um total de 60 publicações. A Alemanha contribuiu com aproximadamente 10% das publicações da amostra, com 52 artigos.

Os 486 artigos da amostra tiveram um total de 2.298 autores, destes mais de 90% publicaram apenas um artigo, 59 autores publicaram dois, 16 autores tiveram três publicações, entre eles Joern Loetsch teve quatro, Rodrigo-David Salas-Rueda seis e Ricardo-Adan Salas-Rueda teve o maior número de publicações com nove artigos. Este resultado, assim como o de Costa e Cunha (2015) e Guimarães e Bezerra (2019), corrobora com a lei de Lotka, que prevê o fato de poucos autores produzirem muito e muitos autores produzirem pouco (COSTA; CUNHA, 2015; GUEDES, BORSCHIVER, 2005). Também é possível traçar um paralelo entre os fenômenos da gestão de dados denominados *small science* e *big science*. De acordo com Sayão e Sales (2019), a *small science* é caracterizada pela grande quantidade de pesquisadores autônomos publicando dados de seus estudos, entretanto, em um somatório de publicações maior do que as *big sciences,* compostas por grandes equipes de pesquisadores produzindo dados de forma estruturada e padronizada para permitir o compartilhamento e reuso dos mesmos.

Dentre os periódicos científicos com mais publicações sobre gestão de dados estão: *IEEE Access*, com cobertura de assuntos relacionados à eletrônica, com 15 artigos; *PLOS ONE*, cobrindo Ciência e Medicina e *Sensors*, que aborda Ciência e Tecnologia com sete artigos cada. As áreas do conhecimento com maiores números de publicações foram a Ciência da Computação (113 artigos), seguida pela Engenharia (65) e pela Ciência da Informação e Biblioteconomia (55). Esses resultados apontam para diversidade de áreas que estudam a temática de gestão de dados, o que já havia sido indicado por Costa e Cunha (2015) e Zhang e Eichmann-Kalwara (2019). É possível refletir sobre o fato da produção e do compartilhamento de dados perpassar as pesquisas que são realizadas nas diversas áreas do conhecimento, tendo o dado como insumo básico. Nesse contexto, o estudo da gestão de dados de pesquisa ser realizado por diferentes áreas do conhecimento é justificado.

A visualização da rede de coocorrência de palavras-chave por meio do *VOSviewer* permitiu identificar os termos indexadores de artigos mais utilizados e suas relações com outros termos, por meio de *clusters* (agrupamentos de termos relacionados). Os três maiores *clusters* foram sobre *data science*, *data management* e *big data.* A *data science* estava ligada a duas vertentes, a educacional e a tecnológica, abordando temáticas como aprendizado de máquina, mineração de dados e tecnologias voltadas à inovação. *Data management* compreendeu termos relacionados à operacionalização da gestão de dados, como, ferramentas, *softwares* e sistemas para compartilhamento, tratamento e organização de dados que permitam sua preservação a longo prazo e a representação da informação digital por meio de metadados. E o fenômeno do *big data* foi relacionado a três indexadores: gestão, análise de dados e inteligência artificial. A gestão permite a estruturação do grande volume de dados em meio digital, as análises sobre estes dados permitem realizar inferências que podem ou não gerar novas teorias e, a inteligência artificial funciona como o instrumento necessário para o acesso e processamento em alta velocidade de grandes volumes de dados.

Em relação ao conjunto de 486 artigos investigados, aproximadamente 60% (303 artigos) tiveram uma ou mais citações, acumulando-se 2.471 citações no período de 2017 a 2021. Ocorreu o aumento no número de citações à medida que houve crescimento da quantidade de publicações, o que pode ser justificado pelo crescente interesse sobre a temática gestão de dados, conforme já apontado por Szigeti e Wheeler (2011), Costa e Cunha (2015), Kirkpatrick (2015), Guimarães e Bezerra (2019) e Cunha e Costa (2020). É possível inferir que haverá uma tendência de crescimento sobre o tema nos próximos anos, pois 2021 apresentou 159 citações,

enquanto em 2017 foram realizadas apenas 40. Este resultado é corroborado por estudos sobre citações que afirmam que estas levam mais tempo, aproximandamente 2 a 5 anos, para acumularem uma quantidade significativa para análise (LIN; FENNER, 2013; PRIEM; PIWOWAR; HEMMINGER, 2012; THELWALL; WILSON, 2015). Os cinco artigos mais citados foram: Chen, I-Min A. *et al*. (2019) *IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes* com 214 citações; Zhang, Dong *et al*. (2020) *PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies* com 130; Karpatne, Anuj *et al*. (2017) *Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data* com 120; Lowndes, Julia S. Stewart *et al*. (2017) *Our path to better science in less time using open data science tools* com 77, e Wang, Lizhe *et al*. (2018) *pipsCloud: High performance cloud computing for remote sensing big data management and processing* com 61.

Conclusões

A presente pesquisa teve como objetivo levantar as características e analisar o desempenho da produção científica sobre gestão de dados de pesquisa por meio de técnicas bibliométricas e da visualização destes com o uso do *software VOSviewer*. Foram analisados metadados de 486 artigos sobre gestão de dados de pesquisa recuperados na *Web of Science*.

As publicações foram analisadas usando-se indicadores bibliométricos. Foi observado um crescimento das publicações entre 2017 e 2020, havendo menor quantidade de publicações em 2021 por ser um ano em andamento. O caráter multidisciplinar da discussão sobre a gestão de dados foi comprovado pelas publicações ocorrerem na Ciência da Computação (113 artigos), Engenharia (65) e Ciência da Informação e Biblioteconomia (55), assim como demonstrado por estudos correlatos (CUNHA; COSTA, 2020; FAGHMOUS *et al*., 2014; SIEBRA *et al*., 2015). Os países com maior número de publicações foram os EUA (32,92%), seguido por Inglaterra (12,34%) e Alemanha (10,7%). Na amostra 2220 autores tiveram apenas uma publicação, os autores com mais artigos publicados foram Ricardo-Adan Salas-Rueda (9 artigos), Rodrigo-David Salas-Rueda (6) e Joern Loetsch (4). O artigo mais citado foi de Chen, I-Min A. *et al*. (2019) *IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes* com 214 citações. Os termos indexadores mais encontrados nos artigos foram *Data Science*, *data management* e *Big Data*, o que indica possíveis caminhos de atuação no atual paradigma da ciência orientada a dados.

Os resultados do estudo bibliométrico permitiram delinear o avanço das produções sobre gestão de dados de pesquisa, o aumento da quantidade de publicações ao longo dos anos pesquisados, além da visibilidade gerada por autores, periódicos e organizações. A realização deste estudo permitiu compreender como está ocorrendo o desempenho da literatura sobre gestão de dados e seus padrões de pesquisa.

A limitação desta pesquisa foi o estudo com enfoque no desempenho acadêmico das publicações, desconsiderando o impacto social destas, que poderá ser aferido por meio de indicadores altmétricos. Sugere-se como pesquisas futuras, a ampliação das bases de dados utilizadas no estudo bibliométrico e a combinação deste com indicadores altmétricos.

Referências bibliográficas

BOERES, S.; CUNHA, M. B. Competências para a preservação e curadoria digitais. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 14, n. 3, p. 426-449, set./dez. 2016.

CODATO, A. Utilizando citações para além do fator de impacto: uma alternativa para determinar topografias científicas. **SciELO 20 Years Repository**, p. 1- 19, set. 2018.

COSTA, M. M.; CUNHA, M. B. A literatura internacional sobre e-Science nas bases de dados LISA e LISTA. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 20, n. 44, p. 127-144, set./dez., 2015.

CUNHA, M. B.; COSTA, M. M. Fontes de informação sobre gestão de dados de pesquisa. **Informação & Sociedade**: Estudos, João Pessoa, v. 30, n. 4, p. 1-59, out./dez. 2020.

FAGHMOUS, J. H. *et al*. Theory-Guided Data Science for Climate Change. **IEEE Computer Society**, 2014.

GUEDES, V. L. S.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. *In*: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO, 6., jun. 2005, Salvador. **Anais** […]. Salvador: UFBA, 2005. 18 p.

GUIMARÃES, A. J. R.; BEZERRA, C. A. Gestão de dados: uma abordagem bibliométrica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 24, n. 4, p. 171-186, out./dez. 2019.

KIRKPATRICK, K. Putting the Data Science into Journalism. **Communications of the ACM**, v. 58, n. 5, p. 15-17, 2015.

LIN, J.; FENNER, M. The many faces of article-level metrics. **Bulletin of the American Society for Information Science and Technology**, v. 39, n. 4, p. 27-30, 2013.

LIMA, S. H. de O.; LEOCÁDIO, Á. L. Mapeando a produção científica internacional sobre inovação aberta. **Revista Brasileira de Gestão e Inovação**, Caxias do Sul, v. 5, n. 2, p. 181- 208, jan./abr. 2018.

PALLUDETO, A. W. A.; FELIPINI, A. R. Panorama da literatura sobre a financeirização (1992-2017): uma abordagem bibliométrica. **Economia e Sociedade**, Campinas, v. 28, n. 2, p. 313-337, maio/ago. 2019.

PRIEM, J.; PIWOWAR, H. A.; HEMMINGER, B. M. **Altmetrics in the Wild**: Using Social Media to Explore Scholarly Impact. [*S. n. d.*], 2012.

SACARDO, M.; HAYASHI, M. C. P. I. Bibliometria e Epistemologia: balanços iniciais da produção científica em educação física na interface com a educação. *In*: HAYASHI, M. C. P. I. *et al*. (org.). **Bibliometria e cientometria**: estudos temáticos. São Carlos: Pedro & João, 2013. p. 85-99.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade**: Estudos, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Tendências da Pesquisa Brasileira em Ciência da Informação**, João Pessoa, v. 6, n. 1., 2013, 23 p.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90–115, maio/ago. 2016.

SAYÃO, L. F.; SALES, L. F. A ciência invisível: os dados da cauda longa da pesquisa científica. *In*: DIAS, G. A.; OLIVEIRA, B. M. J. F. (org.). **Dados científicos**: perspectivas e desafios. João Pessoa: Ed. UFPB, 2019. p. 33-52.

SIEBRA, S. A. *et al*. Curadoria digital: um termo interdisciplinar. *In*: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais** […]. Salvador: UFBA, 2016. 17 p.

SOUSA, E. S. *et al*. Mapeamento da produção científica internacional sobre intenção empreendedora. **Revista Gestão e Secretariado**, São Paulo, v. 10, n. 3, set./dez. 2019, p. 114-139.

SZIGETI, K.; WHEELER, K. Essential Readings in e-Science. **Issues in Science and Technology Librarianship**, Wint. 2011.

THELWALL, M.; WILSON, P. Mendeley readership altmetrics for medical articles: An analysis of 45 fields. **Journal of the Association for Information Science and Technology**, v. 67, p. 1962-1972. 2015.

ZHANG, L.; EICHMANN-KALWARA, N. Mapping the Scholarly Literature Found in Scopus on Research Data Management: A Bibliometric and Data Visualization Approach. **Journal of Librarianship and Scholarly Communication**, v. 7, general issue, 2019.