



Em Questão
ISSN: 1807-8893
ISSN: 1808-5245
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

Qualidade de dados em gestão de dados de pesquisa: um estudo bibliométrico

Piccolo, Daiane Marcela; Tadini, Antonio Victor Wolf; Teixeira, Heytor Diniz; Botega, Leonardo Castro; Sant'Ana, Ricardo César Gonçalves; Santarem Segundo, José Eduardo; Alves, Rachel Cristina Vesu

Qualidade de dados em gestão de dados de pesquisa: um estudo bibliométrico

Em Questão, vol. 28, núm. 1, 2022

Universidade Federal do Rio Grande do Sul, Brasil

Disponível em: <https://www.redalyc.org/articulo.oa?id=465669358007>

DOI: <https://doi.org/10.19132/1808-5245281.159-184>

Qualidade de dados em gestão de dados de pesquisa: um estudo bibliométrico

Data quality in research data management: a bibliometric study

Daiane Marcela Piccolo 1

Universidade Estadual Paulista, Brasil

daiane.piccolo@unesp.br

DOI: <https://doi.org/10.19132/1808-5245281.159-184>

Redalyc: [https://www.redalyc.org/articulo.oa?](https://www.redalyc.org/articulo.oa?id=465669358007)

id=465669358007

Antonio Victor Wolf Tadini 2

Universidade Estadual Paulista, Brasil

antonio.vw.tadini@unesp.br

Heytor Diniz Teixeira 3

Universidade Estadual Paulista, Brasil

hd.teixeira@unesp.br

Leonardo Castro Botega 4

Universidade Estadual Paulista, Brasil

leonardo.botega@unesp.br

Ricardo César Gonçalves Sant'Ana 5

Universidade Estadual Paulista, Brasil

ricardo.santana@unesp.br

José Eduardo Santarem Segundo 6

Universidade de São Paulo, Brasil

santarem@usp.br

Rachel Cristina Vesu Alves 7

Universidade Estadual Paulista, Brasil

rachel.vesu@unesp.br

Recepción: 18 Marzo 2021

Aprobación: 02 Julio 2021

RESUMO:

NOTAS DE AUTOR

- 1 Doutoranda; Universidade Estadual Paulista, Marília, SP, Brasil; daiane.piccolo@unesp.br
- 2 Mestrando; Universidade Estadual Paulista, Marília, SP, Brasil; antonio.vw.tadini@unesp.br
- 3 Mestrando; Universidade Estadual Paulista, Marília, SP, Brasil; hd.teixeira@unesp.br
- 4 Doutor; Universidade Estadual Paulista, Marília, SP, Brasil; leonardo.botega@unesp.br
- 5 Doutor; Universidade Estadual Paulista, Marília, SP, Brasil; ricardo.santana@unesp.br
- 6 Doutor; Universidade de São Paulo, Ribeirão Preto, SP, Brasil; santarem@usp.br
- 7 Doutora; Universidade Estadual Paulista, Marília, SP, Brasil; rachel.vesu@unesp.br

A gestão dos dados de pesquisa é reconhecida pela comunidade científica como parte importante das boas práticas de pesquisa. Desta maneira, acredita-se que os mesmos devem estar sempre disponíveis para acesso e reuso. Neste contexto, a curadoria e a qualidade de dados são entendidas como elementos estratégicos. Este trabalho tem como objetivo caracterizar e especificar a produção científica existente sobre o tema “qualidade de dados em gestão de dados de pesquisa”, por meio da aferição de indicadores bibliométricos. Em termos metodológicos, esta pesquisa possui natureza quantitativa e qualitativa, é de tipo exploratória quanto a seus objetivos e utiliza-se das bases de dados Web of Science e Scopus para a composição do corpus do estudo bibliométrico. Como resultado, identificou-se, a partir de um corpus de 77 artigos, um período de publicações relevantes entre os anos de 1984 e 2020, sendo o ano de 2019 aquele com mais trabalhos publicados. Adicionalmente, 7 veículos de publicação apresentaram mais de um artigo no tópico pesquisado, sendo os Estados Unidos o país com mais trabalhos publicados, totalizando 34. A área da Ciência da Computação foi a que mais produziu nesse tema e constitui uma tendência em sua interdisciplinaridade com as ciências biológicas, sociais aplicadas e da saúde. Finalmente, conclui-se que, a partir da consciência de que a qualidade de dados é um parâmetro relativo, a implementação de serviços de gestão de dados de pesquisa deve passar por preparação, com foco no atendimento a requisitos como os concernentes ao domínio e aos usos pretendidos.

PALAVRAS-CHAVE: Gestão de dados de pesquisa, Qualidade de dados, Dados de pesquisa, Dados científicos, Curadoria digital.

ABSTRACT:

Research data management is recognized by the scientific community as an important part of best practices in research, so that these data should be available for access and reuse. Within the context of research data management, data curation and data quality are understood as strategic elements. This work aims to characterize and specify the existing scientific production on the theme “data quality in research data management” through the measurement of bibliometric indicators. In methodological terms, this research has a quantitative and qualitative nature, is exploratory in its objectives and uses the Web of Science and Scopus databases to compose the corpus of the bibliometric study. As a result, it was identified from a corpus of 77 articles a period of relevant publications between the years 1984 and 2020, being 2019 the year with more published works. Additionally, 7 publication vehicles presented more than one publication for the researched topic, being the United States the country with more published papers, totaling 34 articles. The area of Computer Science was the one that produced the most on this topic and constitutes a trend in its interdisciplinarity with biological, applied social and health sciences. Finally, we conclude that, based on the awareness that data quality is a relative parameter, the implementation of research data management services must go through preparation, focusing on meeting requirements such as those concerning the domain and the intended uses.

KEYWORDS: Research data management, Data quality, Research data, Scientific data, Digital curation.

1 INTRODUÇÃO

Junto às discussões sobre o acesso aberto às pesquisas científicas, surge o debate a respeito do acesso aberto também aos dados que possibilitaram a construção dessas pesquisas (BRASE; FARQUHAR, 2011; SAYÃO; SALES, 2012; SILVA; SANTAREM SEGUNDO; SILVA, 2018).

A expansão do conceito de acesso livre, incorporando agora coleções de dados de pesquisa, vem se consolidando amparada por várias ações cultivadas no próprio seio das comunidades científicas, que reconhecem esses estoques de informação como uma parte do patrimônio da ciência universal e um pilar imprescindível para o seu avanço. O acesso aos dados de pesquisa torna-se, portanto, um imperativo para a ciência com reflexos globais, dado que os pesquisadores trabalham em cooperação internacional e os dados são criados, compartilhados e acessados globalmente (BRASE; FARQUHAR, 2011[1] *apud* SAYÃO; SALES, 2012, p. 181).

O entendimento hoje é de que esses dados, ditos de pesquisa, são parte integrante do registro acadêmico e devem estar disponíveis para reuso, pois vêm sendo reconhecidos pela comunidade científica como parte essencial das boas práticas de pesquisa (SILVA; SANTAREM SEGUNDO; SILVA, 2018).

Sales e Sayão (2019), ao propor uma representação hierárquica para os conceitos presentes no domínio dos dados de pesquisa, examinam definições do conceito de dado de pesquisa a fim de postularem uma definição própria.

Dado de pesquisa é todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito como necessário para validar os resultados da pesquisa pela comunidade científica. (SALES; SAYÃO, 2019, p. 36).

Tanto Sales e Sayão (2019) quanto Silva, Santarem Segundo e Silva (2018) referem-se a Borgman (2010) para reafirmar a dificuldade envolvida na definição do conceito. Borgman (2010) atribui essa dificuldade ao entendimento de que a existência do dado de pesquisa enquanto tal é relativa à interpretação de alguém que dessa maneira o enxergue.

A Figura 1 consiste em um mapa conceitual cuja finalidade é fornecer uma orientação introdutória acerca das relações de significado existentes entre conceitos que são estratégicos na pesquisa. O mapa foi elaborado de modo a representar os conceitos de acordo com as concepções teóricas em que se fundamenta este artigo, as quais podem ser verificadas em Marín-Arraiza, Puerta-Díaz e Vidotti (2019), Sales *et al.* (2019), Silva, Santarem Segundo e Silva (2018) e Tartarotti, Dal'Evedove e Fujita (2019).

Figura 1 - Mapa conceitual: qualidade de dados em gestão de dados de pesquisa

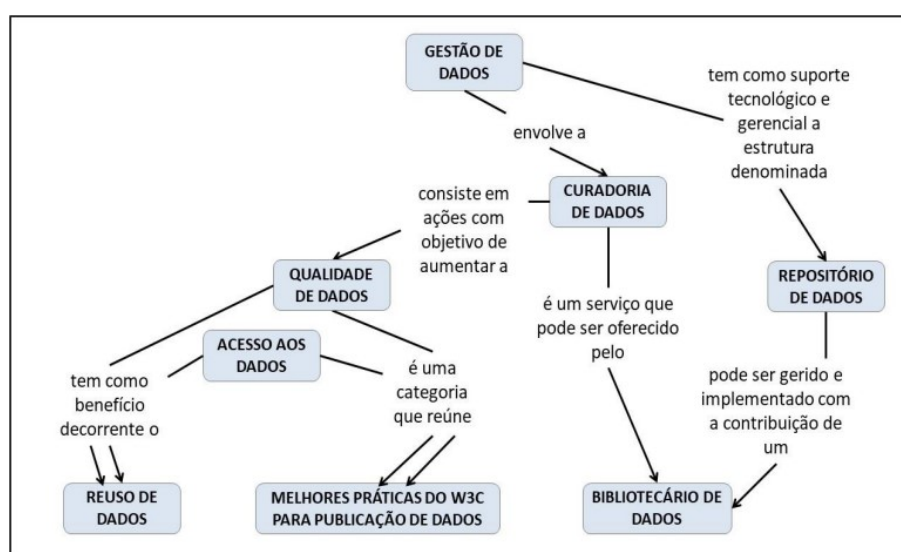


FIGURA 1

Mapa conceitual qualidade de dados em gestão de dados de pesquisa

Fonte: Elaborado pelos autores.

Note-se que as questões de acesso, a despeito de serem fundamentais, não bastam: elas são tão importantes quanto o foco na viabilização do reuso dos dados – enquanto benefício decorrente, entre outros fatores, do próprio acesso aos dados, bem como da qualidade dos dados (LÓSCIO; BURLE; CALEGARI, 2017; SILVA; SANTAREM SEGUNDO; SILVA, 2018).

O objetivo deste trabalho é especificar e caracterizar a produção científica existente sobre o tema “qualidade de dados em gestão de dados de pesquisa”. Para tanto, a aferição de indicadores bibliométricos objetiva, nesta pesquisa, fornecer subsídios para a exploração da intersecção em que consiste a incidência da qualidade de dados na gestão de dados de pesquisa.

O estudo bibliométrico que ora se apresenta almeja, desse modo, prover “informações sobre a estrutura do conhecimento e sua comunicação”, de modo a evidenciar características e comportamentos da comunicação científica referente a esse tópico específico do conhecimento, tendo em vista a dimensão coletiva da construção do conhecimento humano (BUFREM; PRATES, 2005, p. 23).

1.1 Acesso e reuso de dados de pesquisa

Com relação à viabilização do acesso a dados de pesquisa de um modo geral, GABRIEL JUNIOR *et al.* (2019, p. 90) afirmam que:

Os benefícios mais evidentes são a possibilidade da reprodução ou da verificação da pesquisa; a disponibilização dos resultados de pesquisas financiadas com fundos públicos; a continuidade das pesquisas e dos questionamentos a respeito dos dados existentes e, conseqüentemente, a viabilização de avanços no estado da pesquisa e na inovação.

Tenopir *et al.* (2015) mostram que o acesso aberto aos dados de pesquisa, do ponto de vista do pesquisador, proporciona visibilidade; para as instituições de ensino superior, melhora o monitoramento e avaliação das atividades científicas, além de proporcionar prestígio e visibilidade; para as entidades financiadoras, justifica o investimento, cria transparência e evita duplicação de financiamento; e, especificamente na pesquisa, pode beneficiar a comunidade científica global.

Contudo, conforme mencionado anteriormente, a mera disponibilização dos dados de pesquisa não é o bastante. É nesse sentido que Wilkinson *et al.* (2016) apontam para o gerenciamento dos dados como um canal estratégico para o conhecimento, a descoberta e a inovação, e subsequente integração e reuso de dados e conhecimento pela comunidade após a publicação dos dados. Silva, Santarem Segundo e Silva (2018) esclarecem que os princípios FAIR (*Findable, Accessible, Interoperable, Re-usable*), referidos originalmente por Wilkinson *et al.* (2016), orientam na elaboração do plano de gestão e publicação de dados e aprimoram a capacidade das máquinas em utilizar e encontrar os dados, além de apoiar a reutilização desses dados por indivíduos.

Wilkinson *et al.* (2016, p. 4) afirmam que os princípios FAIR, apesar de serem elementos relacionados, são independentes e separáveis, e atuam como guia para auxiliar os atores envolvidos a avaliar suas escolhas de implementação. Com base em Lóscio, Burle e Calegari (2017), é possível vislumbrar a relação existente entre os princípios segundo os quais os dados devem ser “encontráveis” (*findable*) e “reusáveis” (*re-usable*) por meio da seguinte declaração: as questões de acesso aos dados ocupam uma das 13 categorias de melhores práticas do W3C para publicação de dados. Essa categoria reúne sozinha 10 do total de 35 práticas elencadas; em comum, todas as 35 práticas, e conseqüentemente as 13 categorias, possuem como benefício decorrente o aprimoramento das condições de reuso.

1.2 Qualidade de dados e os dados de pesquisa

Ainda considerando Lóscio, Burle e Calegari (2017), é possível destacar mais uma dessas 13 categorias de melhores práticas do W3C: a que envolve questões de qualidade de dados.

Melo, Botega e Santarem Segundo (2017, p. 83) entendem – em consonância com o conceito de “*fitness for use*” postulado por Illari e Floridi (2014) – que a ideia de qualidade envolve “medidas para que o produto oferecido esteja de acordo com o que se espera dele, podendo este ser uma informação, um dado, um serviço ou um processo”. Os autores salientam também que os problemas de qualidade nos dados existem quando eles não representam adequadamente os componentes da realidade do domínio em que se inscrevem, contribuindo negativamente para a execução das atividades deles dependentes.

Santos e Sant’Ana (2013, p. 205), para definir o que é um dado, ressaltam que ele é necessariamente relacionado a determinado contexto, cujo detalhamento, mesmo que não esteja explícito, “[...] deverá estar disponível de modo implícito ao utilizador, permitindo, portanto, sua plena interpretação”. A qualidade de dados, em decorrência disso, não pode ser tomada como um parâmetro absoluto, e sim relativo aos requisitos do domínio e dos usos pretendidos para eles (BATINI *et al.*, 2009; MELO; BOTEGA; SANTAREM SEGUNDO, 2017; OLIVER; HARVEY, 2010; SAYÃO; SALES, 2015). Sem essa contextualização, não é

possível afirmar taxativamente em que medida (sob quais métricas) uma porção específica de dados é dotada de qualidade.

É importante considerar que a qualidade de dados é um conceito de sentido delimitado tradicionalmente na literatura científica, de modo que constitui um campo consolidado do conhecimento, que possui arcabouço teórico próprio. Algumas referências que corroboram para essa consolidação podem ser destacadas, como Batini *et al.* (2009), Batini e Scannapieco (2016), Hacid *et al.* (2019), Illari e Floridi (2014), Laudon (1986), entre outras.

A qualidade de dados, de um modo geral, é compreendida em dimensões (de qualidade de dados). A respeito disso, os autores abaixo mencionados apontam que

[...] existem dimensões mais utilizadas, porém não há um padrão estabelecido de dimensões; cada domínio utiliza dimensões que atendam aos seus requisitos específicos. Dentre as mais utilizadas na literatura constam: completude, precisão, relevância e consistência. (MELO; BOTEGA; SANTAREM SEGUNDO, 2017, p. 84).

Devido ao fato de a qualidade de dados não ser um parâmetro absoluto, mas relativo aos requisitos do domínio e dos usos das atividades pretendidas, não é apropriado tentar estabelecer uma metodologia aplicável, com profundidade satisfatória, irrestritamente a quaisquer situações. Desse modo, dimensões funcionam como aspectos pelos quais a qualidade de dados pode ser avaliada de maneira adequada, dentro de um domínio.

Isso pressupõe que, para determinado domínio, sejam definidas essas dimensões, o que pode ser concretizado no âmbito de uma metodologia de avaliação construída com foco nesse contexto específico – a exemplo do que se pode verificar em Melo, Botega e Santarem Segundo (2017), que estabelecem uma metodologia de avaliação de qualidade de dados para dados publicados no contexto do *Linked Data*. Após a definição das dimensões, parte-se para a definição de métricas, sobrepostas às dimensões, para que seja possível aferir, em termos quantitativos, a qualidade dos dados avaliados no que se refere a determinada dimensão. E, então, é possível a obtenção de um panorama quantitativo considerando todas as dimensões definidas para o contexto em que o conjunto de dados avaliado se inscreve (BATINI *et al.*, 2009).

Sant’Ana (2019) concebe a qualidade (assim como a privacidade, a integração, a legislação, a disseminação e a preservação) como um fator transversal que permeia todas as fases do ciclo de vida dos dados (coleta, armazenamento, recuperação e descarte – conforme apresentado pelo autor). No que se refere, especificamente, à gestão de dados de pesquisa, não é diferente: é essencial que se tenha a noção de ciclo de vida desses dados.

Tendo isso em vista, Sayão e Sales (2015) estabelecem uma distinção entre os conceitos de garantia de qualidade e de controle de qualidade: a garantia de qualidade se refere aos processos aplicados antes e durante a coleta dos dados, enquanto o controle de qualidade congrega os processos posteriores à coleta (como “limpeza de dados” e decisões sobre dados ausentes e valores estimados).

Assim como Lóscio, Burle e Calegari (2017), Oliver e Harvey (2010) também posicionam a manutenção da qualidade de dados como uma das estratégias para que os dados sejam “reusáveis” (*re-usable*). Desse modo, emerge a curadoria dos dados de pesquisa, inscrita na gestão de dados de pesquisa e também entendida como “[...] ações mais dinâmicas e contundentes sobre os dados [...], que visam adicionar valor aos dados [...]” (SAYÃO; SALES, 2016, p. 96).

A curadoria digital, em resumo, assegura a sustentabilidade dos dados para o futuro, não deixando, entretanto, de conferir valor imediato a eles para os seus criadores e para os seus usuários. Os recursos estratégicos, metodológicos e as tecnologias envolvidas nas práticas da curadoria digital facilitam o acesso persistente a dados digitais confiáveis **por meio da melhoria da qualidade desses dados**, do seu contexto de pesquisa e da checagem de autenticidade. (SAYÃO; SALES, 2012, p. 185, grifo nosso).

Mais recentemente e de modo mais objetivo, Sales *et al.* (2019, p. 308) definem curadoria de dados como a “manutenção, preservação e agregação de valor a dados de pesquisa durante o seu ciclo de vida”. Desse modo,

sendo a curadoria de dados parte da gestão de dados de pesquisa, é necessário discutir sobre os repositórios de dados, estruturas que dão suporte e constituem um ambiente propício para a gestão de dados de pesquisa.

1.3 Repositório de dados de pesquisa

O termo “gestão de dados” pode ser definido como: “conjunto de atividades gerenciais e tecnológicas, apoiado por políticas gerais e específicas destinadas a garantir: arquivamento, curadoria, preservação e oferta de acesso contínuo aos dados de pesquisa” (SALES *et al.*, 2019, p. 308).

Dá-se o nome de repositório de dados de pesquisa à ferramenta utilizada para dar suporte tecnológico e gerencial às referidas atividades. Sayão e Sales (2015, p. 82) definem repositório de dados de pesquisa como uma “estrutura tecnológica e gerencial que permite que pesquisadores depositem seus dados de pesquisa para armazenamento e amplo acesso”. O repositório está no centro de um arcabouço tecnológico e gerencial que compreende todo o ciclo de vida dos dados para apoiar a execução dos processos envolvidos na gestão de dados de pesquisa (SAYÃO; SALES, 2016).

Vale observar que, apesar de a definição de Sayão e Sales (2015) destacar o armazenamento e o acesso como finalidades, os repositórios de dados de pesquisa dão suporte também à produção, ao uso, ao reuso e ao compartilhamento dos conjuntos de dados gerados durante as várias etapas do processo da pesquisa científica (RICE; SOUTHALL, 2016; TARTAROTTI; DAL'EVEDOVE; FUJITA, 2019).

Os repositórios de dados de pesquisa também podem ser descritos como “plataformas colaborativas de gestão de dados de pesquisa” (SAYÃO; SALES, 2018, p. 82). Os autores destacam a necessidade de que

[...] os repositórios de dados estejam permeados por políticas organizacionais, condições legais e éticas, processos administrativos, sustentabilidade financeira e temporal e disponham de um elenco de serviços voltados para a sua comunidade-alvo, criando um ambiente multifacetado de gestão de dados e de colaboração entre os pesquisadores. (SAYÃO; SALES, 2018, p. 82).

O repositório de dados de pesquisa é, dessa forma, um sistema digital que integra diversas funções, e “[...] tem como perspectiva oferecer um ambiente dinâmico e flexível – principalmente pela natureza heterogênea dos dados [...]” (SAYÃO; SALES, 2016, p. 99). Desse modo, evidencia-se a intersecção desses repositórios com o tema dos ambientes informacionais digitais, objeto da Arquitetura da Informação e da Ciência da Informação (MARÍN-ARRAIZA; PUERTA-DÍAZ; VIDOTTI, 2019; TORINO; ROA-MARTÍNEZ; VIDOTTI, 2020).

Corroborando para a perspectiva de aderência dos repositórios à Ciência da Informação e às práticas bibliotecárias, verifica-se que o profissional identificado como bibliotecário de dados deve contribuir para a implantação e a gestão dos repositórios de dados de pesquisa, definindo o escopo, escolhendo o esquema de metadados; gerenciando o acesso; revisando a qualidade dos dados; planejando a preservação digital; promovendo repositórios digitais confiáveis; e possibilitando a interoperabilidade. O bibliotecário de dados pode, então, ser entendido como um agente otimizador perante os desafios da gestão dos dados de pesquisa (MARÍN-ARRAIZA; PUERTA-DÍAZ; VIDOTTI, 2019; TORINO; ROA-MARTÍNEZ; VIDOTTI, 2020; RICE; SOUTHALL, 2016).

2 METODOLOGIA

A metodologia utilizada neste trabalho é, sob o ponto de vista de sua abordagem, de natureza quantitativa e qualitativa. Adicionalmente, de acordo com seus objetivos, esta pesquisa possui caráter exploratório.

Tendo em vista o objetivo deste trabalho, a escolha das bases de dados se deu pelos critérios de relevância internacional da base de dados, sua capacidade de possibilitar a geração de um panorama global sobre o tema

pesquisado e o histórico da base de dados de possuir estudos de alta qualidade indexados. Portanto, foram utilizadas para o desenvolvimento desta pesquisa as bases Web of Science e Scopus.

Em se tratando de um estudo bibliométrico, vale compreender que os procedimentos metodológicos aplicados são concernentes a “organização, classificação e avaliação quantitativa sobre padrões de publicação, sujeitas a cálculos matemáticos e estatísticos”, e que a exploração de bases de dados on-line nesse contexto consiste em um procedimento “[...] não somente para analisar documentos ou fatos, mas também para traçar as tendências e o desenvolvimento da sociedade, das disciplinas científicas e das áreas de produção e consumo” (BUFREM; PRATES, 2005, p. 23).

Para a realização das análises bibliométricas, foram observados os seguintes indicadores, dentro do tópico “qualidade de dados em gestão de dados de pesquisa”:

- a) categorização e distribuição quantitativa de publicações para o tópico em função das categorias “estudo teórico” e “estudo empírico”;
- b) distribuição quantitativa de publicações para o tópico em função do ano de publicação;
- c) distribuição quantitativa de publicações para o tópico em função dos veículos de publicação mais produtivos;
- d) distribuição quantitativa de publicações para o tópico em função dos principais países de publicação;
- e) distribuição quantitativa de publicações para o tópico em função das áreas do conhecimento mais produtivas, na Web of Science e na Scopus.

Para instrumentalização da coleta de dados, a pesquisa baseou-se no protocolo de revisões sistemáticas de Kitchenham (2004). No Quadro 1 estão dispostas as strings para as buscas nas bases de dados.

Quadro 1 - Strings de busca nas bases de dados

QUADRO 1
Strings de busca nas bases de dados

Base de dados	String de busca
Web of Science	TI=("data quality") AND TS=("scientific data management" OR "research data management " OR "scientific data curation" OR "research data curation" OR "data curation" OR "scientific data" OR "research data" OR "data repository")
Scopus	TITLE("data quality") AND TITLE-ABS-KEY("scientific data management" OR "research data management" OR "scientific data curation" OR "research data curation" OR "data curation" OR "scientific data" OR "research data" OR "data repository")

Fonte: Elaborado pelos autores.

As buscas foram realizadas entre os dias 15 de junho e 15 de julho de 2020, de modo que retornaram apenas estudos publicados até esse período. A pesquisa na base de dados Web of Science foi realizada pela busca avançada, em que foi utilizada a “Coleção principal da Web of Science” (índices SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC). Já na Scopus foi selecionado somente o campo “Título, resumo, palavras-chave” para a busca com a *string*. Desse modo, foram recuperados com as *strings*, considerando as duas bases de dados, 129 trabalhos no total, sendo 46 na Web of Science e 83 na Scopus.

Para triagem dos estudos, definiram-se duas etapas: (1) leitura de título, resumo e palavras-chave, a fim de verificar o contexto do estudo e a possível existência de trabalhos duplicados, etapa em que foram descartados 41 estudos; (2) leitura técnica subsidiária do conteúdo dos textos (introdução, objetivos e resultados), para verificar se os estudos atendiam aos critérios de inclusão, etapa em que foram descartados 11 estudos.

Desse modo, foram incluídos no corpus os estudos que tratavam do tópico estabelecido (qualidade de dados em gestão de dados de pesquisa), conforme os critérios de inclusão, dentre eles: (a) os trabalhos que avaliavam outros estudos sobre a temática; (b) os que forneciam procedimentos para a avaliação da qualidade de dados; (c) os que avaliavam estruturas, metodologias ou softwares já aplicados para a avaliação da qualidade de dados.

Para obtenção, tabulação e representação gráfica dos resultados concernentes às análises bibliométricas, foram utilizados o Excel, como ferramenta para edição de planilhas, bem como ferramentas disponibilizadas nas próprias bases de dados, destinadas à realização de filtragens e análises.

3 RESULTADOS

A seleção decorrente da avaliação dos documentos recuperados para o tópico “qualidade de dados em gestão de dados de pesquisa” gerou um corpus composto por 77 publicações. O Quadro 2 apresenta o resultado da coleta, em ordem cronológica decrescente, identificando as publicações selecionadas por meio das colunas “ID” (identificação dos trabalhos), “Título do artigo”, “Autores”, “Ano” e a coluna “Categoria” com os semi-identificadores “T” e “E”, em que “T” identifica estudo teórico e “E” identifica estudo empírico.

Quadro 2 - Publicações selecionadas para as análises bibliométricas

Publicações selecionadas para as análises bibliométricas

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

Fonte: Elaborado pelos autores.

O Quadro 2 mostra que o corpus é constituído por 29 estudos teóricos e 48 empíricos. Na Figura 2, é possível visualizar a distribuição dos trabalhos conforme o ano de publicação.

Figura 2 - Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função do ano de publicação

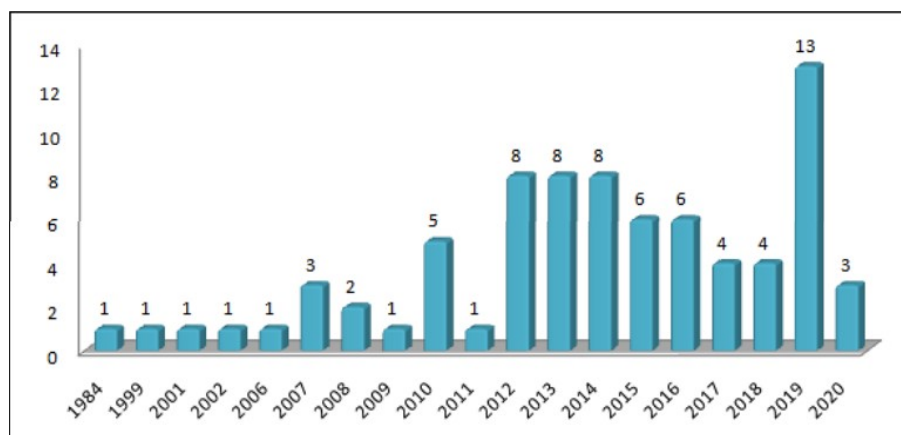


FIGURA 2

Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função do ano de publicação

Fonte: Elaborado pelos autores.

Nota-se, pela Figura 2, que as publicações científicas identificadas estão distribuídas entre os anos de 1984 e 2020. Os dados revelam uma curva instável ao longo do tempo. O ano de 2019 destacou-se pelo maior número de trabalhos publicados, com um total de 13 publicações, das quais 12 foram categorizadas como referentes a estudos empíricos.

Nota-se, ademais, que 31% das publicações se concentram, aproximadamente, entre os anos de 2012 e 2014. Por meio de exame dos títulos, resumos e palavras-chave das publicações do corpus, observou-se que esse dado pode estar relacionado à expansão das pesquisas sobre curadoria, sob tal terminologia, bem como à consolidação do Digital Curation Centre (DCC). É notável, nesse sentido, que as fases 1, 2 e 3 do DCC tenham sido finalizadas respectivamente em 2007, 2010 e 2013 (DIGITAL CURATION CENTRE, 2021).

A análise subsequente evidencia os veículos de publicação indexados na Web of Science e na Scopus que se destacaram como os mais produtivos para o tópico da pesquisa, conforme exposto na Figura 3.

Figura 3 – Distribuição quantitativa de publicações em função dos veículos de publicação mais produtivos para o tópico qualidade de dados em gestão de dados de pesquisa

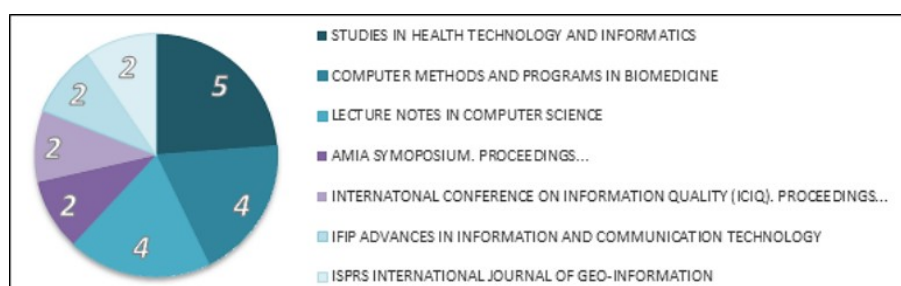


FIGURA 3

– Distribuição quantitativa de publicações em função dos veículos de publicação mais produtivos para o tópico qualidade de dados em gestão de dados de pesquisa

Fonte: Elaborado pelos autores.

A Figura 4 apresenta uma análise do volume das publicações de cada país. Em consonância com os filtros e a análise das bases de dados utilizadas, o país de publicação é referente à nacionalidade atribuída aos autores, de modo que, se a publicação é assinada por três indivíduos com nacionalidades distintas, por exemplo, os três respectivos países pontuam uma vez com base nessa publicação.

Figura 4 - Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função dos principais países de publicação

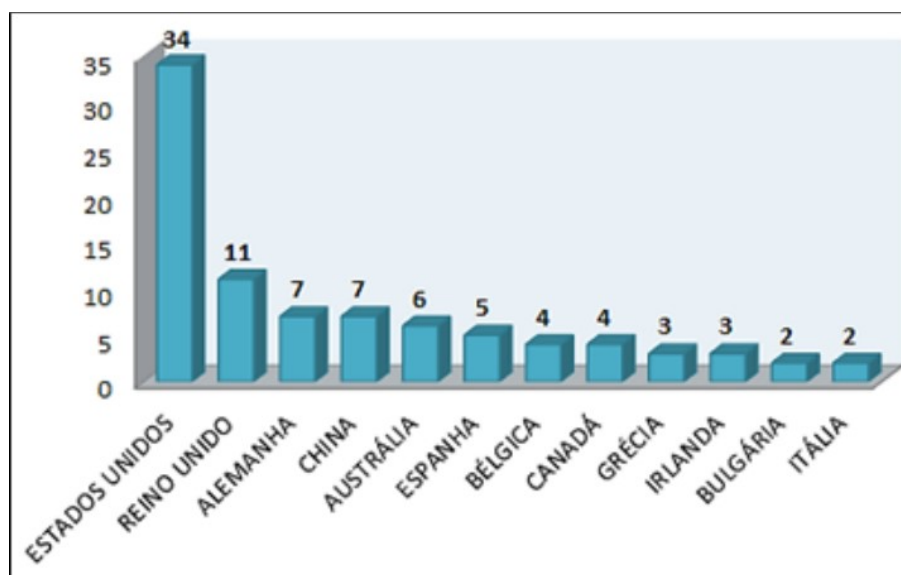


FIGURA 4

Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função dos principais países de publicação

Fonte: Elaborado pelos autores.

Os países que publicaram (1 artigo cada), mas não estão na Figura 4 são: Arábia Saudita, Brasil, Coreia do Sul, Dinamarca, Eslovênia, França, Gana, Hungria, Israel, Japão, Luxemburgo, Noruega, Nova Zelândia, Polônia, Portugal, Singapura e Suíça.

Posteriormente, foi realizada a análise das áreas do conhecimento que, para o tópico, mais tiveram registros identificados nas bases de dados. A Figura 5 mostra os gráficos das áreas mais produtivas em cada base. Devido ao fato de as bases apresentarem diferentes políticas de classificação temática, julgou-se necessária a elaboração de dois gráficos separadamente, dispensando uma eventual compilação dos dados provenientes das duas bases.

Figura 5 - Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função das áreas do conhecimento mais produtivas: Web of Science e Scopus

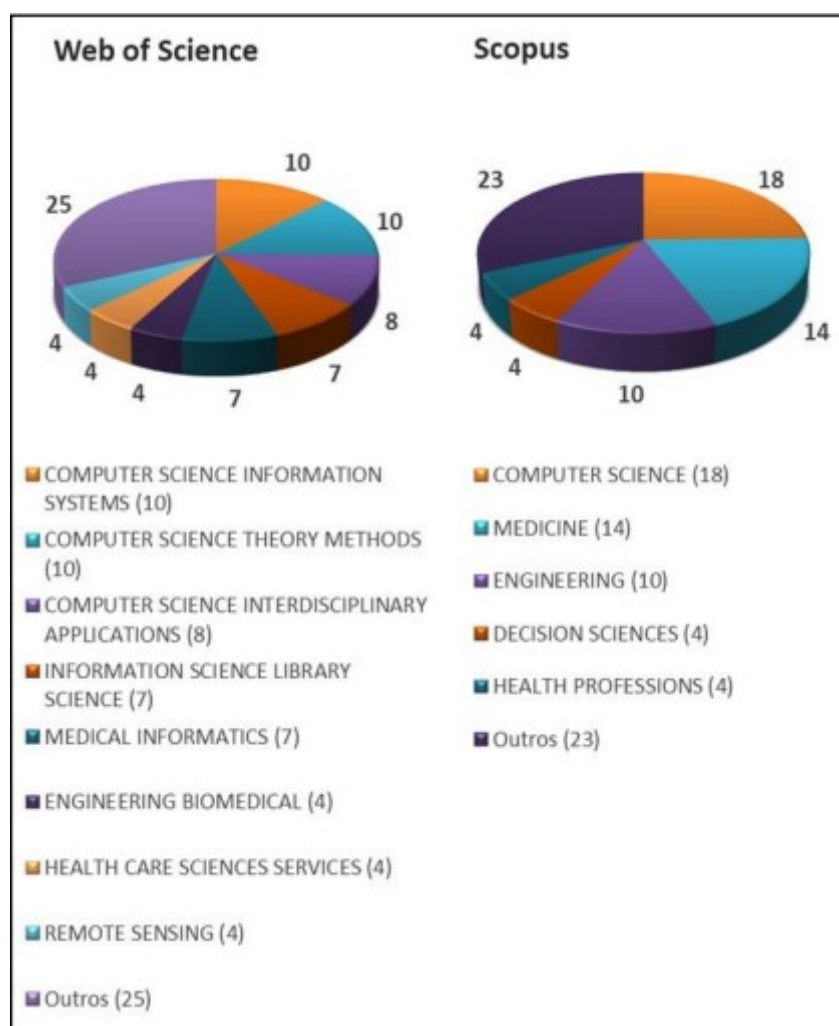


FIGURA 5

Distribuição quantitativa de publicações sobre qualidade de dados em gestão de dados de pesquisa em função das áreas do conhecimento mais produtivas Web of Science e Scopus

Fonte: Elaborado pelos autores.

A partir dos gráficos da Figura 5, é notável um predomínio dos estudos desenvolvidos, em linhas gerais, no âmbito da Ciência da Computação. Desse modo, tal traço se reflete nas análises a seguir.

Partindo da observação de que o ano de 2019 se destaca com a maior quantidade de estudos (13), a análise de tal amostra permite a percepção de uma tendência qualitativa nas produções científicas para o tema. Nela, nota-se a presença marcante de uma relação de interdisciplinaridade que reside na investigação de métodos e aplicações da Ciência da Computação para instrumentalizar a gestão de dados de pesquisa no âmbito das ciências biológicas e da saúde. Considerando esse subconjunto de estudos e a categorização temática da Web of Science, destacam-se subáreas como a Informática Médica, a Engenharia Biomédica e a de Serviços em Saúde.

Rajan *et al.* (2019) explicam sobre a heterogeneidade das fontes no domínio da pesquisa biomédica, de um modo geral, e como isso constitui um problema, uma vez que as práticas nesse contexto estão cada vez mais dependentes da utilização secundária dos dados existentes, a exemplo dos prontuários eletrônicos do paciente. Isso, então, impacta na estrutura de avaliação da qualidade de dados nesse domínio, considerando suas dimensões e métricas. No estudo, os autores desenham, desenvolvem e implementam uma ferramenta para avaliação de qualidade de dados e caracterização de dados em repositórios de dados de saúde.

É possível também tecer uma análise sobre a produção científica concernente à Ciência da Informação, ao se considerar, conforme sustentado anteriormente, que o bibliotecário pode atuar na gestão de dados de pesquisa, o que se soma ao fato de que não existem parâmetros absolutos para a determinação da qualidade de dados, pois tal aferição é necessariamente relativa a domínios específicos, de modo que é necessário se adaptar ao domínio de atuação para melhor atender o usuário.

A partir desses pressupostos, buscou-se verificar se algum dos 77 estudos selecionados para as análises bibliométricas deste trabalho poderia embasar a atuação do bibliotecário com vistas à qualidade de dados em gestão de dados de pesquisa. Adotou-se como estratégia a análise dos 7 estudos categorizados pela Web of Science como sendo de Biblioteconomia e Ciência da Informação. Os estudos estão identificados no Quadro 3.

Quadro 3 - Subconjunto de publicações da categoria Biblioteconomia e Ciência da Informação da Web of Science

QUADRO 3

Subconjunto de publicações da categoria Biblioteconomia e Ciência da Informação da Web of Science

ID	Título do artigo	Autores	Ano
25	Data governance, data literacy and the management of data quality	Koltay, T.	2016
33	Research project tasks, data, and perceptions of data quality in a condensed matter physics community	Stvilia, B. <i>et al.</i>	2015
34	Genomics data curation roles, skills and perception of data quality	Huang, H.; Joergensen, C.; Stvilia, B.	2015
42	Data quality issues and content analysis for research data repositories: the case of Dryad	Rousidis, D. <i>et al.</i>	2014
53	Dynamic data maintenance for quality data, quality research	Ozmen-Ertekin, D.; Ozbay, K.	2012
54	Prioritization of data quality dimensions and skills requirements in genome annotation work	Huang, H. <i>et al.</i>	2012
70	Know thy sensor: trust, data quality, and data integrity in scientific digital libraries	Wallis, J. C. <i>et al.</i>	2007

Fonte: Elaborado pelos autores.

Com base na quantidade de publicações de cada autor e na quantidade de citações fornecida pela Web of Science para cada publicação, é possível destacar como principais autores: Besiki Stvilia, Hong Huang e Corinne Joergensen. Para esse subconjunto, o único país com mais de um estudo são os Estados Unidos: são 5 estudos, nos quais estão inseridos os trabalhos dos autores mencionados, que, por sua vez, concentram-se mais especificamente no estado da Flórida.

Ainda considerando o subconjunto de 7 estudos em Biblioteconomia e Ciência da Informação, 4 deles são categorizados pela Web of Science como pertinentes à Ciência da Computação, dos quais 3 são os que se destacam pelo número de citações na Web of Science: Wallis *et al.* (2007), 26 citações; Huang *et al.* (2012), 14 citações; e Stvilia *et al.* (2015), 10 citações.

Outro estudo em Biblioteconomia e Ciência da Informação que merece destaque, não apenas pelo número de citações na Web of Science (17), mas também por ser o mais recente entre os 7, é Koltay (2016), artigo publicado no IFLA Journal. Nele, o autor aborda questões de qualidade de dados pela ótica da governança de dados (*data governance*), e conclui que, apesar de o conhecimento sobre governança de dados vir recebendo mais atenção em ambientes empresariais, e apesar de os bibliotecários já dominarem

algumas das competências a ele relacionadas, é fundamental sua compreensão e aplicação para beneficiamento dos serviços de gestão de dados de pesquisa. Especialmente porque, como demonstra o autor, técnicas de governança de dados são aplicáveis a todos os níveis presentes em tais serviços.

4 CONSIDERAÇÕES FINAIS

A realização da pesquisa que culminou neste trabalho possibilitou, primeiramente, a realização de um mapeamento teórico a respeito da incidência da qualidade de dados sobre a curadoria e a gestão de dados de pesquisa, cujas atividades são apoiadas pelos repositórios de dados.

A partir dos resultados das análises bibliométricas, este trabalho explicita um panorama geral da produção científica existente sobre qualidade de dados em gestão de dados de pesquisa. Dentro disso, foram possíveis apontamentos, tais como a identificação de uma tendência qualitativa nas produções científicas no tema, localizada na interdisciplinaridade entre a Ciência da Computação e as ciências biológicas e da saúde. Para pesquisas futuras, vislumbra-se a realização de estudos com características ainda mais qualitativas a partir dos resultados apresentados, tanto os referentes ao corpus de publicações ora selecionado e avaliado, quanto os relativos aos dados bibliométricos aferidos.

Assim como o que se apresenta, trabalhos que desenvolvam aspectos envolvidos na gestão de dados de pesquisa contribuem para a Ciência na medida em que fortalecem a cultura da Ciência Aberta e dos princípios FAIR. Nessa perspectiva, são muitos os benefícios, desde os que decorrem diretamente do acesso aberto a dados – bem como da qualidade desses dados – à possibilidade, por meio de seu reuso, de reprodução e verificação das pesquisas, e mesmo de potencialização da visibilidade de pesquisadores e instituições.

Trabalhos como este são importantes por tratarem de um tema em franca expansão, que tem um caminho ainda a enfrentar para sua ampla concretização; mesmo no cenário internacional, mas principalmente no brasileiro.

Aos profissionais da informação – e especificamente aos bibliotecários, este trabalho pode indicar caminhos para a oferta de serviços de gestão de dados de pesquisa. A partir da consciência de que a qualidade de dados é um parâmetro relativo, a implementação de serviços dessa natureza deve passar por preparação, com foco no atendimento a requisitos como os concernentes ao domínio de atuação – que eventualmente pode estar contemplado por literatura específica para qualidade de dados em gestão de dados de pesquisa no referido domínio –, bem como aos usos pretendidos para os dados.

Financiamento

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS

- BATINI, C. *et al.* Methodologies for data quality assessment and improvement. **ACM Computing Surveys**, Nova York, v. 41, n. 3, p. 1-52, 2009.
- BATINI, C.; SCANNAPIECO, M. **Data and information quality: dimensions, principles and techniques**. [S. l.]: Springer, 2016.
- BORGMAN, C. L Research data: who will share what, with whom, when and why?. **RatSWD Working Paper Series**, Berlim, v. 161, n. 10, 2010.
- BRASE, J; FARQUHAR, A. Access to research data. **D-Lib Magazine**, [s. l.], v. 17, n. 1/2, 2011.
- BUFREM, L.; PRATES, Y. O saber científico registrado e as práticas de mensuração da informação. **Ciência da Informação**, Brasília, v. 34, n. 2, p. 9-25, 2005.
- DIGITAL CURATION CENTRE. **History of the DCC**. DCC, 2021.

- GABRIEL JUNIOR, R. F. *et al.* Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. **Ciência da Informação**, Brasília, v. 48, n. 3 (Supl.), p. 87-101, 2019.
- HACID, H. *et al.* **Data quality and trust in big data**. [S. l.]: Springer, 2019.
- HUANG, H. *et al.* Prioritization of data quality dimensions and skills requirements in genome annotation work. **Journal of the American Society for Information Science and Technology**, [Silver Spring], v. 63, n. 1, p. 195-207, 2012.
- ILLARI, P.; FLORIDI, L. Information quality, data and philosophy. In: FLORIDI, L.; ILLARI, P. (Eds.). **The philosophy of information quality**. Cham: Springer, 2014. p. 5-23.
- KITCHENHAM, B. Procedures for performing systematic reviews. **Keele University Technical Report**, Keele, v. 33, p. 1-26, jul. 2004.
- KOLTAY, T. Data governance, data literacy and the management of data quality. **IFLA Journal**, [s.l.], v. 42, n. 4, p. 303-312, 2016.
- LAUDON, K. C. Data quality and due process in large interorganizational record systems. **Communications of the ACM**, Nova York, v. 29, n. 1, p. 4-11, 1986.
- LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web best practices**. W3C, 2017.
- MARÍN-ARRAIZA, P.; PUERTA-DÍAZ, M.; VIDOTTI, S. A. B. G. Gestión de datos de investigación y bibliotecas: preservando los nuevos bienes científicos. **Hypertext.net**, Barcelona, n. 19, p. 13-31, 2019.
- MELO, J. O. S. F.; BOTECA, L. C.; SANTAREM SEGUNDO, J. E. Metodologia de avaliação de qualidade para dados conectados. **Informação & Tecnologia**, Marília/João Pessoa, v. 4, n. 2, p. 80-101, 2017.
- OLIVER, G.; HARVEY, D. R. **Digital curation**. Chicago: ALA Neal-Schuman, 2010.
- RAJAN, N. S. *et al.* Towards a content agnostic computable knowledge repository for data quality assessment. **Computer Methods and Programs in Biomedicine**, [s. l.], v. 177, p. 193-201, 2019.
- RICE, R.; SOUTHALL, J. **The data librarian's handbook**. London: Facet Publishing, 2016.
- SALES, L. F. *et al.* Competências dos bibliotecários na gestão dos dados de pesquisa. **Ciência da Informação**, Brasília, v. 48, n. 3 (Supl.), p. 303-313, 2019.
- SALES, L. F.; SAYÃO, L. F. Há futuro para as bibliotecas de pesquisa no ambiente e Science? **Informação & Tecnologia**, Marília/João Pessoa, v. 2, n. 1, p. 30-52, 2015.
- SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Conhecimento em Ação**, Rio de Janeiro, v. 4, n. 1, p. 31-48, 2019.
- SANT'ANA, R. C. G. Campo informacional resultante da interação de ciclos de vida dos dados. In: DIAS, G. A.; OLIVEIRA, B. M. J. F. **Dados científicos: perspectivas e desafios**. João Pessoa: Ed. UFPB, 2019. p. 13-31.
- SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, n. 2, p. 199-209, 2013.
- 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, 2016.
- 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, 2012.
- SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015.
- SAYÃO, L. F.; SALES, L. F. Subsídios para a construção de um modelo de avaliação de sistemas de gestão de dados de pesquisa. **PontodeAcesso**, Salvador, v. 12, n. 3, p. 80-108, 2018.
- SILVA, L. C.; SANTAREM SEGUNDO, J. E.; SILVA, M. F. Princípios FAIR e melhores práticas do Linked Data na publicação de dados de pesquisa. **Informação & Tecnologia (ITEC)**, Marília/João Pessoa, v. 5, n. 2, p. 81-103, 2018.
- STVILIA, B. *et al.* Research project tasks, data, and perceptions of data quality in a condensed matter physics community. **Journal of the Association for Information Science and Technology**, Silver Spring, v. 66, n. 2, p. 246-263, 2015.

- TARTAROTTI, R. C. D.; DAL'EVEDOVE, P. R.; FUJITA, M. S. L. Biblioteconomia de dados em repositórios de pesquisa: perspectivas para a atuação bibliotecária. *Informação & Informação*, Londrina, v. 24, n. 3, p. 207-226, 2019.
- TENOPIR, C. *et al.* Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One*, São Francisco, v. 10, n. 8, p. 1-24, 2015.
- TORINO, E.; ROA-MARTÍNEZ, S. M.; VIDOTTI, S. A. B. G. Dados de pesquisa: disponibilização ou publicação?. *In: SHINTAKU, M.; SALES, L. F.; COSTA, M. (Org.). Repositórios digitais: teoria e prática*. Botucatu: ABEC, 2020. p. 183-201.
- WALLIS, J. C. *et al.* Know thy sensor: trust, data quality, and data integrity in scientific digital libraries. *In: KOVÁCS, L.; FUHR, N.; MEGHINI, C. (Eds.). Research and advanced technology for digital libraries (Lecture Notes in Computer Science, v. 4675)*. Berlin: Springer, 2007. p. 380-391.
- WILKINSON, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, Londres, v. 3, artigo n. 60018, 2016.

Data quality in research data management: a bibliometric study

Abstract: Research data management is recognized by the scientific community as an important part of best practices in research, so that these data should be available for access and reuse. Within the context of research data management, data curation and data quality are understood as strategic elements. This work aims to characterize and specify the existing scientific production on the theme “data quality in research data management” through the measurement of bibliometric indicators. In methodological terms, this research has a quantitative and qualitative nature, is exploratory in its objectives and uses the Web of Science and Scopus databases to compose the corpus of the bibliometric study. As a result, it was identified from a corpus of 77 articles a period of relevant publications between the years 1984 and 2020, being 2019 the year with more published works. Additionally, 7 publication vehicles presented more than one publication for the researched topic, being the United States the country with more published papers, totaling 34 articles. The area of Computer Science was the one that produced the most on this topic and constitutes a trend in its interdisciplinarity with biological, applied social and health sciences. Finally, we conclude that, based on the awareness that data quality is a relative parameter, the implementation of research data management services must go through preparation, focusing on meeting requirements such as those concerning the domain and the intended uses.

Keywords: Research data management; Data quality; Research data; Scientific data; Digital curation

NOTAS

- 1 BRASE, J; FARQUHAR, A. Access to research data. *D-Lib Magazine*, [s.l.], v. 17, n. 1/2, 2011. Apud Sayão e Sales (2012).

INFORMACIÓN ADICIONAL

Declaração de autoria: **Concepção e elaboração do estudo:** Daiane Marcela Piccolo, Antonio Victor Wolf Tadini, Heytor Diniz Teixeira, Leonardo Castro Botega **Coleta de dados:** Daiane Marcela Piccolo, Antonio Victor Wolf Tadini, Heytor Diniz Teixeira **Análise e interpretação de dados:** Daiane Marcela Piccolo, Antonio Victor Wolf Tadini, Heytor Diniz Teixeira, Leonardo Castro Botega, Ricardo César Gonçalves Sant'Ana, José Eduardo Santarem Segundo, Rachel Cristina Vesu Alves **Redação:** Daiane Marcela Piccolo, Antonio Victor Wolf Tadini, Heytor Diniz Teixeira **Revisão crítica do manuscrito:** Leonardo Castro Botega, Ricardo César Gonçalves Sant'Ana, José Eduardo Santarem Segundo, Rachel Cristina Vesu Alves

Como citar:: PICCOLO, Daiane Marcela et al. Qualidade de dados em gestão de dados de pesquisa: um estudo bibliométrico. **Em Questão**, Porto Alegre, v. 28; n. 1, p. 159-184, 2022. DOI: <http://dx.doi.org/10.19132/1808-5245281.159-184>