

Curadoria digital: estudo bibliométrico na Scopus de 2010 a 2020

Byanca Neumann Salerno

Mestre; Universidade Federal do Paraná, Curitiba, PR, Brasil;
byancasalerno@ufpr.br; ORCID: <https://orcid.org/0000-0002-1609-4684>

Paula Carina de Araújo

Doutora; Universidade Federal do Paraná, Curitiba, PR, Brasil;
paulacarina@ufpr.br; ORCID: <https://orcid.org/0000-0003-4608-752X>

Maria do Carmo Duarte Freitas

Doutora; Universidade Federal do Paraná, Curitiba, PR, Brasil;
mcf@ufpr.br; ORCID: <https://orcid.org/0000-0002-7046-6020>

Resumo: A curadoria digital é uma área de pesquisa cuja relevância tem crescido em função do aumento das informações digitais. Este artigo objetiva analisar a produção científica sobre curadoria digital na base Scopus de 2010 a 2020. Para isso, coleta 334 referências de documentos da Scopus e utiliza os softwares Bibliometrix e VOSViewer. Analisa os resultados a partir de indicadores de produção, ligação e citação, como: tipologia dos documentos, evolução temporal, lei de Lotka, autores mais produtivos e seu impacto, lei de Bradford, fontes de publicação com maior impacto, palavras-chaves usadas pelos autores e sua rede de coocorrência e análise de cocitação. Conclui que a curadoria digital é uma área interdisciplinar com amplas possibilidades de pesquisa e oferece um panorama parcial sobre a curadoria digital capaz de auxiliar pesquisadores a terem um ponto de partida acerca do tema.

Palavras-chave: Curadoria digital; Bibliometria; Scopus; Lei de Lotka; Lei de Bradford; Produção científica

1 Introdução

O volume de dados e informações digitais aumentou nos últimos anos devido a sua criação em formato digital. Além disso, a digitalização de dados e informações que antes estavam em formato físico contribui para esse crescimento (BEAGRIE, 2006; ABBOTT, 2008; SIEBRA; BORBA; MIRANDA, 2016).

Esse fenômeno global gera preocupação com o gerenciamento e a preservação desses ativos (BEAGRIE, 2006; PENNOCK, 2007; ABBOTT, 2008; HIGGINS, 2011), o que deu origem à curadoria digital, definida como:

[...] o gerenciamento e preservação de dados digitais a longo prazo. Todas as atividades envolvidas no gerenciamento de dados, desde o planejamento de sua criação, melhores práticas em digitalização e documentação, e garantia de sua disponibilidade e adequação para descoberta e reutilização no futuro, fazem parte da curadoria digital. (ABBOTT, 2008, p. 1, tradução nossa)¹.

A curadoria é entendida como uma prática interdisciplinar (BEAGRIE, 2006; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; SIEBRA; BORBA; MIRANDA, 2016; PIRES; ROCHA, 2020) que beneficia diferentes áreas que lidam com dados e informações digitais. Por isso, é importante conhecer sua evolução em termos de pesquisas científicas publicadas.

A pesquisa científica é um processo contínuo de geração de conhecimento a partir de conhecimentos prévios (CHUEKE; AMATUCCI, 2015). Nesse sentido, estudos voltados para a compreensão de uma área do conhecimento dão subsídios para que novas pesquisas aconteçam, tendo em vista a quantidade de fontes de informação disponíveis, que torna impraticável o esgotamento de um assunto.

A Ciência da Informação visa “investigar os mecanismos de produção, representação, disseminação e avaliação da informação” (CURTY; DELBIANCO, 2020, p. 2). Assim, o campo dos estudos métricos da informação se destaca como uma maneira de avaliar as informações científicas, considerando seu impacto, influência e alcance. Apesar de enquadrados na Ciência da Informação, os estudos métricos da informação são interdisciplinares e recebem diferentes classificações de acordo com o objeto de análise: bibliometria, infometria, cientometria, cibermetria, webometria, patentometria e arquivometria (CURTY; DELBIANCO, 2020).

O foco deste artigo é a bibliometria, considerada por Araújo (2006, p. 12) como uma “técnica quantitativa e estatística de medição dos índices de produção e disseminação do conhecimento científico”. Em sua origem, a bibliometria media monografias, tais como livros de bibliotecas.

Posteriormente, passa a abranger artigos, periódicos, autores e citações (ARAÚJO, 2006; CURTY; DELBIANCO, 2020), permitindo relatar problemas a serem investigados em pesquisas futuras pelo fato de sistematizar essas informações (CHUEKE; AMATUCCI, 2015).

A bibliometria é regida por três leis principais: lei de Lotka, que investiga a produtividade de autores (LOTKA, 1926); lei de Bradford, que avalia a produtividade de periódicos (BRADFORD, 1934); e a lei de Zipf, que analisa a frequência de palavras (ZIPF, 1972). Neste artigo serão utilizadas as duas primeiras.

De acordo com a lei de Lotka, dada uma área do conhecimento, a proporção de autores que publicam apenas um artigo é de 60%, enquanto os 40% restantes representam um núcleo de autores mais produtivos. Assim, a quantidade de autores com “n” publicações é aproximadamente $1/n^2$ daqueles que fazem apenas uma contribuição (LOTKA, 1926). A lei de Bradford considera que, em um conjunto de periódicos, existem três zonas, cada qual com 1/3 do total de artigos, sendo que a primeira zona é considerada a mais devotada ao tema (BRADFORD, 1934).

Além dessas leis, os estudos bibliométricos utilizam outros indicadores, tais como colaboração entre autores, países e instituições e a análise de citações. Dentro da análise de citações, existem diferentes indicadores. A literatura recomenda não usar esses indicadores isoladamente, pois cada um deles apresenta suas limitações. Sendo assim, é oportuno utilizá-los em conjunto para uma avaliação mais justa e precisa (THOMAZ; ASSAD; MOREIRA, 2011).

Neste artigo foi considerado o índice-h (HIRSCH, 2005), que combina quantidade de publicações e de citações. Se um autor possui índice-h igual a 15, isso significa que ele produziu 15 artigos com pelo menos 15 citações (HIRSCH, 2005). Outros indicadores de citação considerados nesta pesquisa foram o índice-g e o índice-m. O primeiro é uma melhoria do índice-h e representa a quantidade de citações de um grupo de artigos com maior impacto (EGGHE, 2006); o segundo permite comparar a carreira científica dos autores, sendo útil enquanto eles são produtivos (HIRSCH, 2005) (Quadro 1).

Quadro 1 – Significado do índice-m conforme o valor

Valor de m	Significado
≈ 1	Caracteriza um cientista de sucesso
≈ 2	Caracteriza cientistas de destaque pertencentes às melhores universidades ou laboratórios de pesquisa
≈ 3	Caracteriza indivíduos únicos

Fonte: Adaptado de Hirsch (2005).

Sob essa perspectiva, este artigo tem como objetivo: analisar a produção científica sobre curadoria digital na base Scopus de 2010 a 2020 por meio de bibliometria.

Estudos bibliométricos similares também apresentam a curadoria digital como objeto de estudo. Santos (2014) analisa o conceito de 2000 a 2013 e utiliza como indicadores: autoria, palavras-chave, ano de publicação, idioma, tipo de documento, países e instituições. A autora coleta os dados nas bases BRAPCI, LISA, LISTA, ACM DL, ERIC, INSPEC, E-Lis e Web of Science. Trivedi (2019) utiliza a base DOAJ para recuperar documentos do *International Journal of Digital Curation* no período de 2006 a 2017 e considera os indicadores: ano de publicação, coautoria e grau de colaboração. Por fim, Guallar, Codina e Abadal (2020) analisam a curadoria de conteúdo, uma das ramificações específicas da curadoria digital. Os autores utilizam a base Scopus, sem delimitação de período, e os indicadores: tipo de documento, autoria e coautoria, ano de publicação, instituições, periódicos, idioma, palavras-chave e citações.

A justificativa para este estudo reside no fato de que a pesquisa de Santos (2014) termina em 2013 e não contempla a base Scopus; Trivedi (2019) limita o escopo a um *journal* específico e Guallar, Codina e Abadal (2020), embora seja um estudo recente, não aborda a curadoria digital (termo amplo), mas sim a curadoria de conteúdo (termo específico).

A próxima seção apresenta a curadoria digital e a trajetória metodológica em sequência. Posteriormente, relatam-se os resultados da análise feita a partir de indicadores bibliométricos. Finaliza-se com as considerações finais.

2 Curadoria digital

No século XXI, houve um aumento da quantidade de tecnologias e de dados e informações digitais e atividades como compras, entretenimento, serviços

públicos e a comunicação entre as pessoas ganharam sua versão online (PENNOCK, 2007). Organizações de diferentes áreas como educação, administração pública, cultura, ciência e negócios estão criando um volume maior de dados digitais, que podem ser originalmente criados nesse formato ou digitalizados (BEAGRIE, 2006; ABBOTT, 2008; SIEBRA; BORBA; MIRANDA, 2016).

Neste sentido, grande parte da base de conhecimentos, dados e ativos intelectuais de instituições, organizações e indivíduos estão, hoje, cada vez mais em formato digital. Incluindo documentos administrativos, prontuários médicos, objetos de aprendizagem, objetos culturais, transações comerciais, bancárias e de cartões de crédito, e-mails, músicas, fotos, vídeos, filmes, notícias, postagens em redes sociais, jogos online, sites, entre outros (SIEBRA; BORBA; MIRANDA, 2016, p. 22).

Esse aumento não se limita apenas às organizações, mas também abrange o âmbito pessoal dos indivíduos (BEAGRIE, 2006; HIGGINS, 2011; SIEBRA; BORBA; MIRANDA, 2016). Portanto, a sociedade produz e consome essas informações e isso altera a maneira como as pessoas e os sistemas sociais se comportam (SAYÃO; SALES, 2016).

Uma das preocupações associadas ao aumento do volume de dados e informações é a sua preservação e continuidade a longo prazo, pois suportes digitais são frágeis e ficam obsoletos com o avanço tecnológico (BEAGRIE, 2006; PENNOCK, 2007; ABBOTT, 2008; HIGGINS, 2011). Isso é uma ameaça não apenas organizacional, mas também pessoal (BEAGRIE, 2006).

Essa abundância de dados pessoais e coleções apresenta inúmeros desafios para os indivíduos, incluindo: como proteger fisicamente esse material, às vezes ao longo de décadas; como proteger a privacidade; como organizar e extrair conhecimento útil desta rica biblioteca de informações e como usá-lo de forma eficaz [...] (BEAGRIE, 2006, p. 12, tradução nossa).²

Nesse contexto, Siebra, Borba e Miranda (2016) assinalam que :

[...] o termo curadoria digital emerge sustentado pelo crescimento exponencial do volume de dados digitais e necessita de uma sólida compreensão de terminologias, que deve ser estabelecida como referencial de discurso e para a precisão do trabalho de curadoria e produção de fontes de informação e conhecimento (SIEBRA; BORBA; MIRANDA, 2016, p. 33).

Para Beagrie (2006) o termo é derivado da curadoria originalmente feita em bibliotecas, museus e nas ciências biológicas, e foi usado pela primeira vez em 2001 no seminário *Digital Curation: digital archives, libraries and e-science*. Por outro lado, Siebra, Borba e Miranda (2016) afirmam que o termo começou a ser utilizado pela Ciência da Informação e pela Ciência da Computação em 2003. Os autores Pires e Rocha (2020) abordam que a origem do termo está relacionada à gestão dos dados de pesquisa e posteriormente se tornou interesse de outras organizações.

Nesse sentido, a curadoria digital é aplicada a diferentes disciplinas e profissões que envolvam dados e informações digitais (ABBOTT, 2008; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; PIRES; ROCHA, 2020) e seu foco é garantir o acesso à informação para quem dela precise (SAYÃO; SALES, 2012).

A curadoria digital é um conceito em construção (SAYÃO; SALES, 2012) e o termo ainda causa confusão por ser polissêmico e utilizado em diferentes acervos, como o educacional, o museológico e o de dados de pesquisa, por exemplo (SIEBRA; BORBA; MIRANDA, 2016). Apesar disso, é consenso que o ponto central da curadoria digital é a preservação dos dados e informações digitais durante todo o seu ciclo de vida para as gerações atuais e futuras (PENNOCK, 2007; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; PIRES; ROCHA, 2020).

A abordagem a partir do ciclo de vida permite que, mesmo com as mudanças tecnológicas e comportamentais, se saiba exatamente de onde vieram os dados e para onde vão (PENNOCK, 2007). Os autores Machado e Vianna (2016) adicionam a agregação de valor aos documentos digitais como outro aspecto a ser considerado na curadoria digital, o que também é abordado por Pires e Rocha (2020), que defendem ser preciso agregar valor à informação para que ela seja confiável.

O valor da informação na curadoria digital está relacionado com as propriedades da informação, ou seja, aquelas características que não podem se perder com as transformações que ocorrerem durante o seu ciclo de vida. Nesse sentido, é necessária a compreensão do ambiente ao qual a informação pertence,

porque as mudanças, sejam tecnológicas ou comportamentais, colocam sua preservação em risco (PIRES; ROCHA, 2020).

Isso significa que, apesar de as Tecnologias Digitais da Informação e Comunicação (TDICs) continuarem operando, os próprios usuários podem deixar de utilizá-la. Além do aspecto comportamental citado, também se consideram outros fatores como os geológicos e climatológicos, por exemplo. Sendo assim, ao pensar em curadoria digital é preciso considerar todos esses aspectos e não apenas os tecnológicos. Como pontuado por Siebra, Borba e Miranda (2016), a curadoria digital também enfrenta desafios devido à manipulação das informações por diversas pessoas e o contexto em que ocorrem; o uso e reuso da informação e os avanços tecnológicos. A seção seguinte apresenta a metodologia deste estudo.

3 Metodologia

Esta pesquisa é descritiva, básica e combina elementos quantitativos e qualitativos (GIL, 2017). A primeira etapa consistiu na coleta de dados na base Scopus no dia 4 de janeiro de 2021. A escolha desta base se deu por três motivos: (1) ser multidisciplinar; (2) não ter sido abrangida, dentro do escopo pretendido, nos estudos similares citados na introdução deste artigo; (3) e ser aceita pelos softwares escolhidos para a análise.

No processo de busca na Scopus, foi utilizada a expressão *digital curation*, entre aspas, com o parâmetro padrão (título, resumo e palavras-chave). Para delimitação, foi considerado o período diacrônico de 2010 a 2020, o que representa 11 anos. Esse período é adequado para pesquisas bibliométricas de acordo com Chueke e Amatucci (2015), que recomendam no mínimo dez anos. Esta primeira etapa resultou em 353 documentos.

Nas opções de refinamento da busca, foram excluídos os documentos no prelo, ou seja, documentos que foram aceitos, mas não foram publicados. Para isso, foi selecionada a opção “*Final*” no filtro “*Publication Stage*”. Além disso, optou-se por excluir documentos categorizados como *editorial* e *conference review*, o que resultou em um *corpus* de 334 documentos. Os dados foram

exportados em formato BibTeX e CSV sem qualquer tipo de tratamento e foram realizadas as seguintes análises (Quadro 2).

Quadro 2 – Síntese das análises

Análise	Software	Parâmetros
Tipologia dos documentos	Bibliometrix	<i>Dataset; Main Information</i>
Publicação anual		<i>Dataset; Annual Scientific Production</i>
Lei de Lotka		<i>Authors; Lotka's law</i>
Produtividade e impacto dos autores		<i>Authors; Author Impact; Seleção em ordem decrescente dos dez mais produtivos</i>
Lei de Bradford		<i>Sources; Bradford's law; seleção das fontes da primeira zona</i>
Impacto das fontes mais citadas		<i>Sources; Source Impact; Total Citation; Table</i>
Nuvem de palavras		<i>Documents; Wordcloud; Author Keywords; palavras-chave = 40; font size = 2.5; font type = Tahoma</i>
Coocorrência de palavras-chave (Network Visualization)	VosViewer	<i>Author keywords; full counting; mín. ocorrências = 7; exclusão do termo "digital curation"; exclusão de termos não conectados; palavras-chave = 20; scale = 0.93; labels = 1.00; lines = 1.00.</i>
Cocitação de autores (Network Visualization)		<i>Co-citation; Cited authors; full counting; thesaurus; mín. ocorrências = 25; autores = 27; scale = 1.00; labels = 0.5; lines = 0.5.</i>

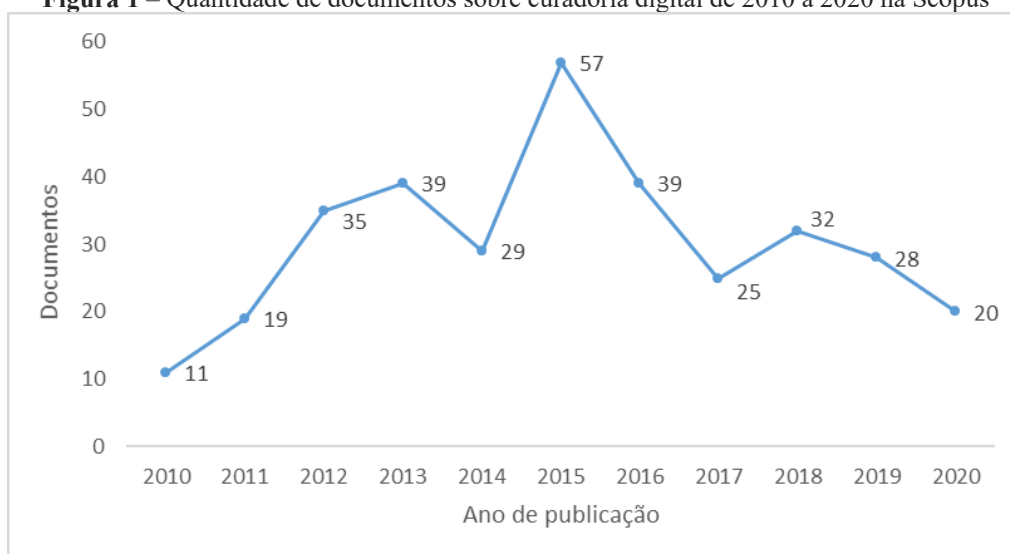
Fonte: Elaborado pelas autoras.

A análise do arquivo BibTeX ocorreu na biblioteca bibliometrix do software R (ARIA; CUCCURULLO, 2017) e a análise do arquivo CSV ocorreu no VOSViewer. Na próxima seção são apresentados os resultados.

4 Apresentação e discussão dos resultados

Dos 334 documentos, havia 163 artigos de periódicos, quatro livros, 14 capítulos de livros, 137 artigos de eventos e 16 artigos de revisão. É relevante saber a evolução por ano, uma vez que Santos (2014) faz esta análise até o ano de 2013 (SANTOS, 2014). Esses documentos distribuem-se conforme Figura 1.

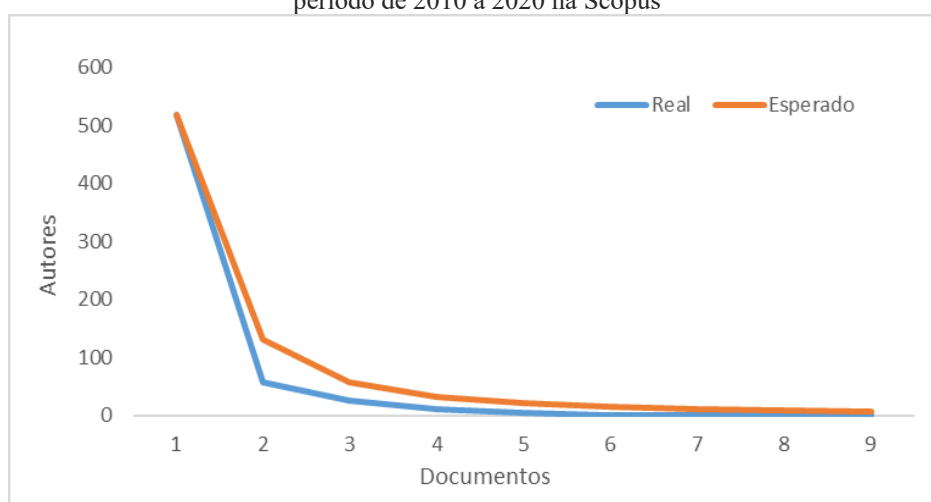
Figura 1 – Quantidade de documentos sobre curadoria digital de 2010 a 2020 na Scopus



Fonte: Dados da pesquisa.

O ano 2015 foi o mais produtivo na última década. Observou-se que dos 57 documentos, 22 são do *CEUR Workshop Proceedings*, dos quais 21 são derivados do *Framing the Digital Curation Curriculum Conference* (DigCurV 2013), um evento específico sobre curadoria digital, o que explica o pico em 2015. Embora reduzida a produção científica a partir desse ano, a taxa de crescimento anual é de 6,16%³ no conjunto de dados. Outra análise importante sob o prisma da produção científica é a lei de Lotka (Figura 2).

Figura 2 – Lei de Lotka: produtividade real e esperada dos autores sobre curadoria digital no período de 2010 a 2020 na Scopus



Fonte: Dados da pesquisa.

Essa lei permite identificar os autores mais produtivos sobre um tema e, pela análise da Figura 2, a proporção real é muito semelhante à proporção esperada, pois poucos autores publicam muitos artigos (LOTKA, 1926). Ressalta-se que os autores mais produtivos não necessariamente correspondem aos autores mais citados, pois um autor pode publicar muitos artigos e receber poucas citações, enquanto outro pode receber muitas citações em poucos artigos. Para avaliar o impacto dos autores mais produtivos, foram considerados aqueles que produziram no mínimo cinco artigos (Tabela 1).

Tabela 1 – Impacto dos autores produtivos sobre curadoria digital de 2010 a 2020 na Scopus

Autor	Índice h	Índice g	Índice m	Citações	Publicações	Início
LEE	2	5	0,17	33	9	2010
MARCIANO	2	4	0,40	16	8	2017
POOLE	4	7	0,40	60	8	2012
TAMMARO	2	3	0,22	15	7	2013
TIBBO	2	3	0,17	14	6	2010
CHRISTODOULAKIS	4	4	0,36	19	5	2011
MAKRIS	4	4	0,36	19	5	2011
MOLLOY	2	4	0,25	19	5	2014
SABHARWAL	2	4	0,29	20	5	2015
SKEVAKIS	4	4	0,36	19	5	2011

Fonte: Dados da pesquisa.

Ao observar somente o indicador de quantidade de publicações, Lee, Marciano, Poole e Tammaro, respectivamente, são os mais produtivos. Dentre eles, Poole e Lee são os mais citados, sendo que Poole possui quase o dobro das citações de Lee e a diferença na quantidade de publicações desses dois autores é de apenas um documento. Ao comparar outros indicadores, Poole apresenta melhores índice-h, índice-g e índice-m em relação a Lee, podendo ser considerado o autor com maior impacto. Marciano e Tammaro possuem os indicadores índice-g e citações próximos e o índice-h igual. Eles diferem no ano de início das publicações e no índice-m, que no caso de Marciano é quase o dobro do de Tammaro, mesmo tendo menor tempo de publicações.

O resultado do índice-m de todos os autores está abaixo de 1, o que permite inferir que as respectivas carreiras acadêmicas estão em estágio de desenvolvimento, não sendo enquadrados ainda como cientistas de sucesso de acordo com Hirsch (2005).

No entanto, este estudo é limitado a 11 anos, o que influencia nesse resultado. Apesar disso, esse índice coloca em perspectiva como está o desenvolvimento da carreira desses autores no período, o que permite compará-los quando outros indicadores são semelhantes.

Nesse sentido, ao comparar os autores que possuem as publicações mais antigas do conjunto (Lee e Tibbo), ambos com início em 2010, verifica-se que, do ponto de vista da carreira científica, seus indicadores são iguais, pois ambos apresentam índice-m igual a 0,17. Lee recebeu mais citações no período em um conjunto de artigos, dado seu índice-g igual a cinco e por possuir três publicações a mais do que Tibbo no período analisado.

Por outro lado, Marciano é o autor mais recente do período, com início em 2017. Considerando o conjunto de dados, ele é um autor promissor, pois apesar de possuir quantidade de citações menor em relação a outros autores, seu índice-h e índice-g são semelhantes aos de Lee e seu índice-m é igual ao de Poole, o que sugere uma carreira bem desenvolvida (HIRSCH, 2005).

Dos cinco principais autores discutidos, Christopher A. Lee e Hellen Tibbo são professores da Escola de Informação e Biblioteconomia Universidade da Carolina do Norte, nos Estados Unidos. Alex H. Poole é professor do Departamento de Ciência da Informação na Faculdade de Computação e Informática da Universidade Drexel, nos Estados Unidos. Richard Marciano é professor do Departamento de Ciência da Computação na Faculdade de Estudos da Informação da Universidade de Maryland nos Estados Unidos. Por fim, Anna Maria Tammaro é professora do Departamento de Engenharia da Informação na Universidade de Parma, na Itália.

Outra análise relevante no que se refere aos indicadores de produção é a lei de Bradford, que permite identificar o núcleo de fontes de publicação mais relevantes dentro de uma temática (Tabela 2).

Tabela 2 – Fontes de publicação sobre curadoria digital no período de 2010 a 2020 na Scopus pertencentes à primeira zona de Bradford

Fontes de publicação	Ranking	Frequência
<i>Ceur Workshop Proceedings</i>	1	28
<i>Lecture Notes in Computer Science</i>	2	16
<i>Communications in Computer and Information Science</i>	3	9
<i>ACM International Conference Proceeding Series</i>	4	8
<i>Proceedings of The ACM/IEEE Joint Conference on Digital Libraries</i>	5	8
<i>Proceedings of The Asist Annual Meeting</i>	6	7
<i>Journal of Documentation</i>	7	6
<i>Proceedings of The Association For Information Science and Technology</i>	8	6
<i>Archival Science</i>	9	5
<i>D-Lib Magazine</i>	10	5
<i>New Review of Information Networking</i>	11	5
<i>Archiving 2014 - Final Program and Proceedings</i>	12	4
<i>College and Undergraduate Libraries</i>	13	4

Fonte: Dados da pesquisa.

Essa análise considerou todas as fontes de publicação pertencentes à primeira zona de acordo com a lei de Bradford, agregando, portanto, periódicos e conferências. Isso significa que essas 13 fontes são responsáveis por cerca de um terço das publicações sobre curadoria digital.

Percebe-se que a curadoria digital vem sendo divulgada principalmente em conferências, pois as conferências são um espaço para apresentação e debate entre pares com *feedback* mais rápido do que em periódicos científicos, ideal para um campo do conhecimento que evolui rapidamente. Assim como na análise de autores, há diferença entre o volume de publicações e a quantidade de citações totais recebidas por essas fontes de publicação (Tabela 3).

Tabela 3 – Indicadores de impacto das fontes mais citadas

Fonte	Índice h	Índice g	Citações	Publicações
<i>Journal of Library Administration</i>	3	4	101	4
<i>Archival Science</i>	3	5	45	5
<i>ACM Transactions on Computer-Human Interaction</i>	1	1	42	1
<i>Lecture Notes in Computer Science</i>	4	5	37	16
<i>IFLA Journal</i>	2	2	37	2
<i>Journal of Documentation</i>	3	5	34	6
<i>New Review of Information Networking</i>	2	5	34	5
<i>Library Hi Tech</i>	3	4	29	4
<i>Conference on Human Factors in Computing Systems</i>	1	1	24	1
<i>Social Media and Society</i>	2	2	23	2

Fonte: Dados da pesquisa.

Os periódicos são as fontes de publicação mais citadas na curadoria digital, sendo o *Journal of Library Administration* o principal, com mais do que o dobro de citações do que o segundo colocado. Este periódico recebeu 101 citações em quatro publicações, diferentemente do *Lecture Notes in Computer*

Science que, apesar de apresentar o maior índice-h (4) e a maior quantidade de publicações (16), possui apenas 37 citações. As fontes que apresentam poucas publicações e muitas citações merecem atenção, pois indicam publicações singulares.

Outrossim, os indicadores de índice-h e índice-g fornecem uma comparação do impacto de cada uma dessas fontes (HIRSCH, 2005; EGGHE, 2006) e quando os dois indicadores são altos, significa que a fonte é uma boa opção para o pesquisador que deseja divulgar seus resultados de pesquisa. Outro ponto a ser discutido é a diferença entre as fontes presentes na primeira zona de Bradford (Tabela 2) e as fontes mais citadas (Tabela 3). Na listagem referente à Bradford, há principalmente conferências, que representam as fontes que mais publicam sobre curadoria digital, enquanto na listagem de fontes mais citadas há predominância de periódicos, embora existam algumas conferências.

Apesar disso, há quatro fontes que se repetem nas duas tabelas: *Lecture Notes in Computer Science*, *Journal of Documentation*, *Archival Science* e *New Review of Information Networking*. Dessas, são conferências: *Lecture Notes in Computer Science*; e periódicos: *Journal of Documentation*, *Archival Science* e *New Review of Information Networking*. Essa é uma informação relevante ao considerar onde publicar um artigo em andamento ou concluído sobre curadoria digital, visto que, além de estarem na primeira zona de Bradford e serem dedicados ao tema, essas fontes também possuem maior impacto dentro do conjunto de dados analisado.

Além do conhecimento dos principais autores e fontes de publicação, o tema foi analisado sob o ponto de vista conceitual por meio de suas palavras-chave. Para esta análise, consideraram-se as 40 palavras-chaves mais frequentes, reunidas em uma nuvem de palavras (Figura 4).

Figura 4 – Nuvem de palavras das 40 palavras-chave mais frequentes nos documentos sobre curadoria digital no período de 2010 a 2020 recuperados na Scopus



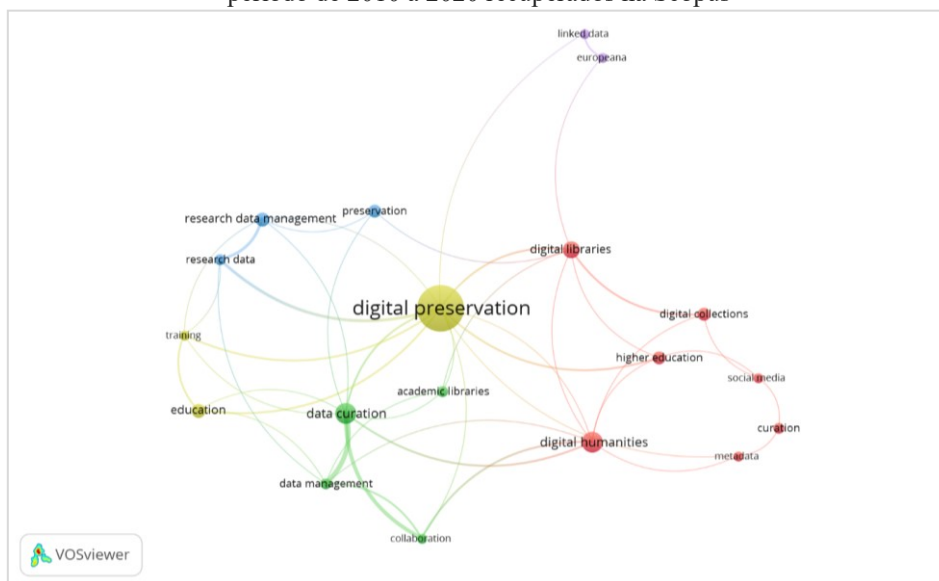
Fonte: Dados da pesquisa.

Nessa nuvem de palavras, o termo *digital curation* foi removido pelo próprio bibliometrix por ser o termo de busca utilizado na Scopus. Assim, a palavra mais frequente nos documentos é *digital preservation*, o que condiz com a preocupação central da curadoria digital (PENNOCK, 2007; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; PIRES; ROCHA, 2020).

Outro ponto que converge com a literatura é a interdisciplinaridade do tema (BEAGRIE, 2006; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; SIEBRA; BORBA; MIRANDA, 2016; PIRES; ROCHA, 2020). Exemplos disso são os termos: *digital humanities*, *education*, *social media*, *cultural heritage*, *curriculum development*, *higher education* e *sustainability*. Apesar dessa relação com outras áreas do conhecimento, é visível a predominância de termos referentes a dados de pesquisa e bibliotecas, fato que foi mencionado pelos autores Pires e Rocha (2020).

A partir dessa primeira interpretação, pode-se organizar esse conjunto de palavras-chaves em *clusters*⁴, de acordo com o seu uso conjunto nos documentos. Esta última análise conceitual foi feita mediante a importação do arquivo CSV no software VOSViewer. Para melhor visualização, foi gerada uma imagem apenas com os termos conectados e considerando no mínimo sete ocorrências, o que resultou em cinco *clusters* (Figura 5).

Figura 5 - Rede de coocorrência de palavras-chave dos documentos sobre curadoria digital no período de 2010 a 2020 recuperados na Scopus



Fonte: Dados da pesquisa.

Assim como na nuvem de palavras-chave (Figura 4), esta primeira rede mostra o termo *digital preservation* ao centro, conectando-se a todos os *clusters*, embora não a todos os termos. Isso reforça a importância da preservação para a curadoria digital, independentemente da área do conhecimento em que ela é aplicada (BEAGRIE, 2006; SAYÃO; SALES, 2012; MACHADO; VIANNA, 2016; SIEBRA; BORBA; MIRANDA, 2016; ROCHA; PIRES, 2020).

No *cluster* amarelo, *education* e *training* ligam-se à *digital preservation*. Isso pode ser explicado pela criação de programas de treinamento e cursos de nível superior em curadoria digital a nível internacional, impulsionados a partir do projeto *Digital Curation Curriculum* (DigCCurr) da Universidade da Carolina do Norte, nos Estados Unidos, que foi objeto de diversas conferências. O propósito do projeto é investigar as necessidades curriculares para o ensino de habilidades em curadoria digital (HIGGINS, 2011).

No *cluster* vermelho, observa-se *higher education* ligando-se a *social media*, *curation*, *metadada*, *digital libraries*, *digital collections* e *digital humanities*. Conforme abordado por Siebra, Borba e Miranda (2016), a educação se aproxima do conceito de curadoria de conteúdo, que se relaciona a seleção, contextualização e compartilhamento de informações disponíveis na

Internet para um público específico. Isso é observado pela presença do termo *social media ligado à higher education*.

Abbott (2008) salienta que o conteúdo produzido no contexto do *e-learning* pode ser reaproveitado em diferentes cenários de aprendizagem e, por isso, é necessária a curadoria e preservação desses recursos. Assim, à medida que as instituições de ensino superior produzem dados e informações digitais, cresce a preocupação com a preservação desses ativos.

Além disso, observa-se a preocupação com a preservação dos dados resultantes de pesquisa e das coleções disponíveis nas bibliotecas, conforme indicado pelos termos *digital collections* e *digital libraries*. Isso é pontuado por Siebra, Borba e Miranda (2016, p. 32), que destacam a necessidade de uma infraestrutura para “para que os dados de pesquisa possam ser preservados e mantidos para uso futuro”.

Destaque-se ainda o *cluster* representado em azul. Os termos *research data* e *research data management* estão ligados ao termo principal do grafo, *digital preservation*. Assim, fica evidente a importância das discussões sobre gestão de dados de pesquisa (dados científicos) no âmbito da curadoria digital. Sayão e Sales (2016, p. 68) destacaram que

o uso e a geração intensiva de dados pelas atividades acadêmicas e de pesquisa criam a necessidade urgente de infraestruturas gerenciais e tecnológicas para que tratem de forma dinâmica o ciclo de vida dos dados (SAYÃO; SALES, 2016, p. 68).

Na parte superior da imagem, há um *cluster* em roxo com dois termos: *européana* e *linked data*. O primeiro refere-se a uma biblioteca virtual com foco na preservação do patrimônio cultural, desenvolvida pela União Europeia. A Europeia “capacita o setor de patrimônio cultural na sua transformação digital por meio do desenvolvimento de conhecimento, ferramentas e políticas para abraçar a mudança digital e encorajar parcerias que fomentam e inovação” (EUROPEANA FOUNDATION, 2021a). O segundo termo do *cluster* em roxo tem relação com a web semântica. Os termos *linked data* (dados ligados) e *linked open data* (dados abertos ligados) foram mencionados pela primeira vez em 2006 em um texto de Tim Berners-Lee. Define-se *linked data* como os

princípios e as melhores práticas para publicação e ligação de dados estruturados na Web (ARAKAKI; SANTOS, 2017).

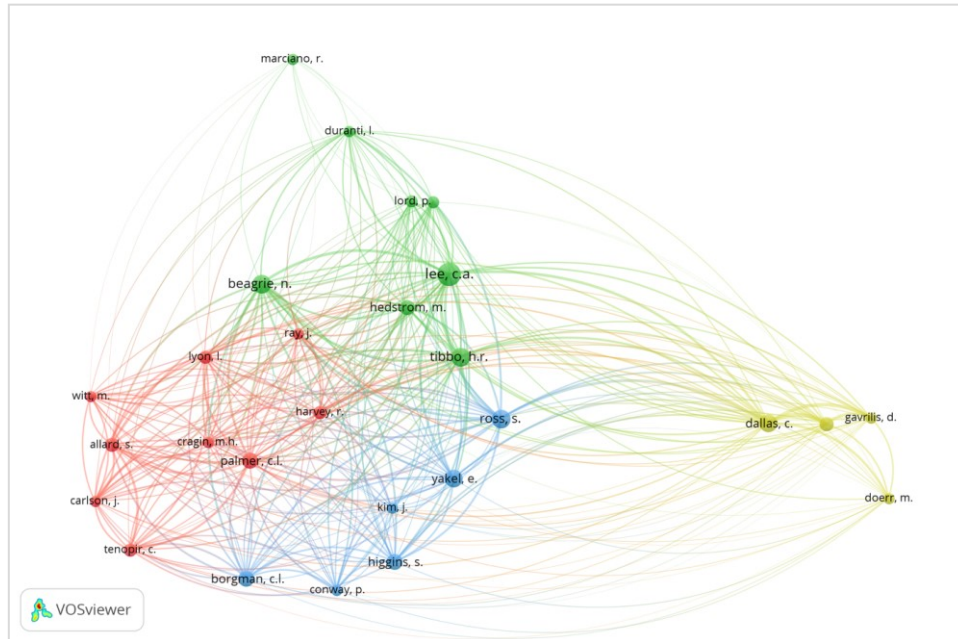
Apresentados os conceitos das palavras-chave do cluster em roxo, é possível inferir que a direta relação entre os termos *européana* e *linked data* justifica-se pelo projeto Europeana de *Linked Open Data*, por meio do qual os metadados de todos os objetos do Portal Europeana são abertos e disponibilizados sob uma licença CC0 Domínio Público sob os termos do *Data Exchange Agreement* (DEA) e podem ser baixados gratuitamente por meio de API (EUROPEANA FOUNDATION, 2021b).

A iniciativa de disponibilização e da preservação dos objetos culturais da Europa, por meio da Europeana, são inovadoras. O projeto Europeana de *Linked Open Data* ultrapassa a barreira do ser humano ao disponibilizar os metadados como um *dataset* do *Linked Open Data* e, dessa forma, “permite que os objetos digitais estejam disponíveis também para acesso por agentes computacionais inteligentes” (CONEGLIAN; SANTARÉM SEGUNDO, 2017, p. 98). Coneglian e Santarém Segundo (2017, p. 98) também reconhecem que ao “organizar a descrição de seus recursos ligando seus objetos a *datasets* públicos do LOD, o *Dataset* da Europeana apresenta-se como um dos principais modelos de informação cultural de acesso público e semanticamente disponível na web”.

Finalizada a análise das palavras-chave, buscou-se, ainda, conhecer os referentes teóricos dos autores do *corpus* que compõe esta pesquisa por meio da análise de cocitação de autor. Acredita-se, assim como Smiraglia (2015), que as citações evidenciam, se certa forma, as relações temáticas e/ou teóricas entre os autores cocitados, o que pode inclusive levar ao reconhecimento de paradigmas dentro de um domínio.

A intensidade da cocitação entre dois autores é determinada pelo número de publicações em que ambos são citados juntos. Para visualizar esta representação, apresenta-se na Figura 6 a rede de cocitação formada por quatro *clusters* compostos de 27 autores cocitados com um mínimo de 25 citações por autor.

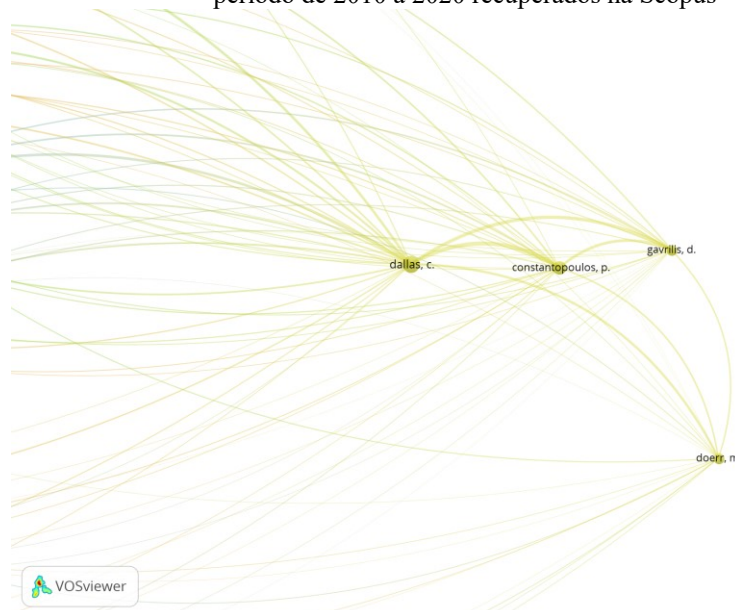
Figura 6 – Rede de cocitação de autores dos documentos sobre curadoria digital no período de 2010 a 2020 recuperados na Scopus



Fonte: Dados da pesquisa.

Os dados obtidos por meio da geração do grafo demonstram que os autores Dallas e Constantopoulos são os mais cocitados da rede. A força de ligação entre eles é de 337. A segunda maior força de ligação é entre Dallas e Ggavrilis, totalizando 224 cocitações (Figura 7).

Figura 7 – Cluster dos autores mais cocitados pelos documentos sobre curadoria digital no período de 2010 a 2020 recuperados na Scopus



Fonte: Dados da pesquisa.

Para complementar esta análise, apresenta-se a Tabela 4 com a indicação dos autores mais citados e a respectiva força total de ligação.

Tabela 4 – Autores mais citados e a força total de ligação

Autores	Citações	Força total de ligação
LEE	113	1960
BEAGRIE	73	1688
ROSS	73	1673
TIBBO	73	1451
DALLAS	68	1692
YAKEL	63	1163
BORGMAN	53	1053
PALMER	52	1463
HIGGINS	52	1073
HEDSTROM	50	999

Fonte: Dados da pesquisa.

O autor mais citado (113) e com a maior força total de ligação (1960) é Lee, o que significa dizer que é o autor mais cocitado com os outros autores da rede. Além de também desempenhar o papel de autor ponte (conforme Figura 6), uma vez que liga os diferentes *clusters*. O segundo autor mais cocitado é Dallas (68) com força total de ligação 1692.

Ao retomar os resultados apresentados na Tabela 1 de autores mais produtivos, percebe-se a coincidência de alguns autores com os dados da Tabela 4 dos mais citados e a força total de ligação. O autor mais produtivo e também o mais cocitado é Lee. Além dele, Tibbo também figura como um dos mais produtivos (Tabela 1) e um dos mais cocitados (Tabela 4), totalizando 73 citações. Os dados relativos à rede de cocitação (Figuras 6 e 7) e aos autores mais citados (Tabela 4) permitem inferir que os autores mais citados podem ser considerados seminais para os estudos de curadoria digital. A análise de cocitação revela a proximidade teórica, epistemológica e/ou metodológica entre esses pares de autores e um estudo futuro com foco na literatura citada pelo domínio pode confirmar essas relações.

5 Considerações finais

A curadoria digital é um conceito interdisciplinar em construção que ganha espaço em diferentes áreas. Sua principal preocupação é a preservação de dados e informações digitais a partir do seu ciclo de vida, em virtude da obsolescência

dos suportes tecnológicos. Nesse contexto, a bibliometria é uma estratégia útil para compreender e ampliar o debate sobre este tema.

Este artigo teve como objetivo analisar a produção científica sobre curadoria digital na base Scopus de 2010 a 2020. Este objetivo foi alcançado com a metodologia e as análises empregadas. Os resultados permitem obter um panorama, ainda que parcial, sobre a curadoria digital. Isso pode auxiliar futuros pesquisadores, bem como aqueles familiarizados com a área a terem um ponto de partida para identificar os principais autores, periódicos, eventos e temas de estudo dentro da curadoria digital.

Apesar disso, esta pesquisa apresenta algumas limitações: a utilização de apenas uma base de dados — Scopus —, o que torna a interpretação dos dados restrita ao seu contexto; a delimitação temporal, que afeta diretamente no resultado dos indicadores, especialmente os de citação; a delimitação dos parâmetros das análises, existindo outras possibilidades a serem exploradas no mesmo conjunto de dados; e delimitação das análises bibliométricas, uma vez que o foco foi a lei de Lotka, a lei de Bradford e alguns indicadores de citação.

Por fim, sugerem-se futuras pesquisas que analisem a dinâmica do tema curadoria digital considerando apenas documentos em acesso aberto, uma vez que este artigo não utilizou este filtro; realizem análises de citação complementares a partir da estrutura intelectual (acoplamento bibliográfico) e social (colaboração) em combinação com análises teóricas e epistemológicas do domínio; listem e discutam, a partir de uma visão crítica, os documentos de maior impacto sobre o tema.

Referências

ABBOTT, D. DCC briefing paper: what is digital curation? Digital Curation Centre: Edinburgh, 2008.

ARAKAKI, F. A.; SANTOS, P. L. V. A. C. Linked data em bibliotecas: iniciativas e tendências. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017. **Anais [...]**... Marília, SP: Unesp, 2017.

ARAÚJO, C. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 11–32, 2006.

ARIA, M.; CUCCURULLO, C. Bibliometrix: an R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, [s. l.], v. 11, n. 4, p. 959–975, 2017.

BEAGRIE, N. Digital curation for science, digital libraries, and individuals. **The International Journal of Digital Curation**, Bath, v. 1, n. 1, p. 3–16, 2006.

BRADFORD, S. C. Sources of information on specific subjects. **Engineering**, [s. l.], v. 137, p. 85–86, 1934.

CHUEKE, G. V.; AMATUCCI, M. O que é bibliometria? Uma introdução ao Fórum. **Internext: Revista Eletrônica de Negócios Internacionais**, São Paulo, v. 10, n. 2, p. 1–5, 2015.

CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Europeana no linked open data: conceitos de web semântica na dimensão aplicada das humanidades digitais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 22, n. 48, p. 88-99, 2017.

CURTY, R. G.; DELBIANCO, N. R. As diferentes metrias dos estudos métricos da informação: evolução epistemológica, inter-relações e representações. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 25, p. 01–21, 2020.

EGGHE, L. Theory and practise of the g-index. **Scientometrics**, Budapest, v. 69, n. 1, p. 131–152, 2006.

EUROPEANA FOUNDATION. **About us**. Den Haag, 2021a. Disponível em: <https://pro.europeana.eu/about-us/mission>. Acesso em 17 maio 2021.

EUROPEANA FOUNDATION. **Linked Open Data**. Den Haag, 2021b. Disponível em: <https://pro.europeana.eu/page/linked-open-data>. Acesso em 17 maio 2021.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2017.

GUALLAR, J.; CODINA, L.; ABADAL, E. La investigación sobre curación de contenidos: análisis de la producción académica. **Ibersid**: revista de sistemas de información y documentación, [s. l.], v. 14, n. 1, p. 13–22, 2020.

HIGGINS, S. Digital curation: the emergence of a new discipline. **The International Journal of Digital Curation**, Bath, v. 6, n. 2, p. 78–88, 2011.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, [s. l.], v. 102, n. 46, p. 16569–16572, 2005.

LOTKA, A. J. The frequency distribution of scientific productivity. **Journal of the Washington Academy of Sciences**, [s. l.], v. 16, n. 12, p. 317–323, 1926.

MACHADO, K. C.; VIANNA, W. B. Curadoria digital e ciência da informação: correlações conceituais relevantes para apropriação da informação. In: XVII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2016. **Anais [...]**. Bahia: ENANCIB, 2016.

PENNOCK, M. Digital curation: a life-cycle approach to managing and preserving usable digital information. **Library & Archives Journal**, [s. l.], v. 1, p. 1–3, 2007.

PIRES, C. de O.; ROCHA, R. P. da. Finalidade e atividades da curadoria digital na perspectiva de sua implantação em uma instituição. **Brazilian Journal of Information Science: Research trends**, [s. l.], v. 14, n. 4, p. 01–31, 2020.

SANTOS, T. N. C. **Curadoria Digital**: o conceito no período de 2000 a 2013. 2014. Dissertação (Mestrado em Ciência da Informação) – Curso de Pós-Graduação em Ciência da Informação, Universidade de Brasília, Brasília, 2014.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ**: novas práticas em informação e conhecimento, [s. l.], v. 5, n. 2, p. 67–71, 2016.

SAYÃO, L. F.; SALES, L. F. Um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade: Estudos**, João Pessoa, v. 22, n. 3, p. 179–191, 2012.

SIEBRA, S. de A.; BORBA, V. da R.; MIRANDA, M. K. F. de O. Curadoria Digital: um termo interdisciplinar. **Informação & Tecnologia**, João Pessoa, v. 3, n. 2, p. 21–38, 2016.

SMIRAGLIA, R. P. **Domain analysis for knowledge organization**: tools for ontology extraction. Waltham, MA: Elsevier, 2015.

THOMAZ, P. G.; ASSAD, R. S.; MOREIRA, L. F. P. Uso do fator de impacto e do índice h para avaliar pesquisadores e publicações. **Arquivos Brasileiros de Cardiologia**, [s. l.], v. 96, n. 2, p. 90–93, 2011.

TRIVEDI, G. Scholarly communication in international journal of digital curation from 2006-2017: a DOAJ based study. **Library Philosophy and Practice**, Lincoln, v. 2866, 2019.

ZIPF, G. K. **Human behavior and the principle of least effort: an introduction to human ecology**. New York: Hafner Publishing Company, 1972.

Digital Curation: bibliometric study at Scopus from 2010 to 2020

Abstract: Digital curation is an area of research whose relevance has grown due to the increase in digital information. This article analyzes the scientific output about digital curation indexed on Scopus database from 2010 to 2020. It collects 334 references from Scopus documents and uses the software Bibliometrix and VOSViewer. It analyses the results related to production, connection and citation indicators, like: documents' typology, temporal evolution, Lotka's law, authors that are more productive and their impact, Bradford's law, sources of publication with more impact, keywords used by the authors and their co-occurrence network, and co-quotation analysis. It concludes that digital curation is an interdisciplinary area with vast research possibilities. The results offer a partial overview of digital curation capable of helping researchers to have a starting point and understand the stage of development of the theme.

Keywords: Digital curation; Bibliometrics; Scopus; Lotka's Law; Bradford's Law; Scientific production

Recebido: 22/03/2021

Aceito: 19/06/2021

Declaração de autoria

Concepção e elaboração do estudo: Byanca Neumann Salerno

Coleta de dados: Byanca Neumann Salerno.

Análise e interpretação de dados: Byanca Neumann Salerno, Paula Carina de Araújo.

Redação: Byanca Neumann Salerno, Paula Carina de Araújo, Maria do Carmo Duarte Freitas.

Revisão crítica do manuscrito: Paula Carina de Araújo, Maria do Carmo Duarte Freitas.

Como citar:

SALERNO, Byanca Neumann. ARAÚJO, Paula Carina de; FREITAS, Maria do Carmo Duarte. Curadoria digital: estudo bibliométrico na Scopus de 2010 a 2020. **Em Questão**, Porto Alegre, v. 28. n. 1, p.185-208, 2022. <https://doi.org/10.19132/1808-5245281.185-208>



¹ No original: “[...] the management and preservation of digital data over the long-term. All activities involved in managing data from planning its creation, best practice in digitisation and documentation, and ensuring its availability and suitability for discovery and re-use in the future are part of digital curation.” (ABBOTT, 2008, p. 1).

² No original: “This abundance of personal data and collections presents numerous challenges to individuals, including: how physically to secure such material sometimes over decades; how to protect privacy; how to organise and extract useful knowledge from this rich library of information and to use it effectively [...]” (BEAGRIE, 2006, p. 12).

³ Taxa calculada automaticamente pelo software Bibliometrix.

⁴ Agrupamento de dados de acordo com características semelhantes.