Artigo Original

Encontros Bibli

AUTOMAÇÃO DA INDEXAÇÃO: EVIDÊNCIAS E TENDÊNCIAS DA PRODUÇÃO CIENTÍFICA INDEXADA NA BRAPCI

Automatic of Indexing: evidence and trends of scientific production indexed in the Brapci

Gustavo Diniz do Nascimento
Universidade Federal da Paraíba, Programa de Pós-graduação em Ciência da Informação, João Pessoa-PB, Brasil
dinizufcg@hotmail.com
https://orcid.org/0000-0002-5130-4149

Gracy Kelli Martins
Universidade Federal da Paraíba, Departamento de Ciência da Informação/ Programa de Pós-graduação em Ciência da Informação, João Pessoa-PB, Brasil
gracykelli@gmail.com
https://orcid.org/0000-0002-1805-9292

Maria Elizabeth Baltar Carneiro de Albuquerque
Universidade Federal da Paraíba, Departamento de Ciência da Informação / Programa de Pós-graduação em Ciência da Informação, João Pessoa-PB, Brasil
ebaltar2007@gmail.com
https://orcid.org/0000-0003-4934-5918

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: As atividades que os profissionais Bibliotecários desempenham estão se adaptando às novas formas de comunicação e às demandas modernas dos usuários. Dentre essas atividades, tem-se a indexação. O presente artigo possui o objetivo de descrever o panorama de produção científica e as tendências de pesquisa sobre os estudos voltados para a indexação automática, semiautomática e auxiliada por computador, indexada na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Brapci), no período de no período de 1972 a 2021.
Método: Caracteriza-se como exploratória e bibliográfica com fins descritivos, uma vez que descreve e detalha a produção científica sobre os temas "indexação automática", "indexação semiautomática" e "indexação auxiliada por máquina" indexada na Brapci.
Resultado: Por meio da análise desse *corpus,* foi possível identificar os pesquisadores que mais produziram sobre a temática em questão, bem como os periódicos que mais tratam da referida temática, além disso, perceberam-se os períodos de tempo mais fecundos à produção de trabalhos relacionados com a automação da indexação e as palavras-chave utilizadas à indexação do corpus analisado, as quais, juntamente com os resumos, sinalizaram as tendências de estudos nessa vertente.
Conclusões: Há uma tendência de estudos voltados à indexação totalmente automática, dando ênfase aos métodos semânticos e não apenas estatísticos, fazendo uso de recursos advindos da Inteligência Artificial. Poucas propostas de indexação semiautomática foram evidenciadas. Propostas promissoras na prática da representação temática da informação de forma automática em curtos espaços de tempo se apresentam hodiernamente.
PALAVRAS-CHAVE: Indexação Automática. Indexação Semiautomática. Indexação auxiliada por Máquina. Representação Temática da Informação. Produção Científica.

ABSTRACT

Objective: The activities that professional Librarians perform are adapting to new forms of communication and the modern demands of users. Among these activities, there is indexing. This article aims to describe the panorama of scientific production and research trends on studies aimed at automatic, semi-automatic and computer-aided indexing, indexed in the Reference Database for Journal Articles in Information Science (Brapci), in the period from 1972 to 2021.
Methods: It is characterized as exploratory and bibliographical with descriptive purposes, since it describes and details the scientific production on the topics "automatic indexing", "semi-automatic indexing" and "machine-aided indexing" indexed in Brapci.
Results: Through the analysis of this corpus, it was possible to identify the researchers who produced the most on the subject in question, as well as the journals that most deal with that subject, in addition, the most fruitful periods of time for the production of related works with the automation of indexing and the keywords used to index the analyzed corpus, which, together with the abstracts, signaled the trends of studies in this area.
Conclusions: There is a tendency for studies aimed at fully automatic indexing, emphasizing semantic and not just statistical methods, making use of resources derived from Artificial Intelligence. Few proposals for semi-automatic indexing

were evidenced. Promising proposals in the practice of thematic representation of information automatically in short spaces of time are presented today.
KEYWORDS: Automatic Indexing. Semiautomatic indexing. Machine-aided indexing. Thematic Representation of Information. Scientific production.

1 INTRODUÇÃO

A forma de produção, consumo e organização da informação tem mudado e, consequentemente, a sociedade vivencia uma crescente produtividade informacional, sobretudo no ambiente digital, que, apoiada pelos avanços provenientes das Tecnologias de Informação e Comunicação, tem resultado em um caos informacional. Diante do excesso de informação das diversas formas de comunicação, os usuários da informação (sejam de bibliotecas, de centros de documentação, dentre outras unidades de informação) hodiernamente estão cada vez mais exigentes. Nesse cenário, as instituições que lidam com a guarda, organização e disseminação de informações invariavelmente vêm se adequando às demandas tecnológicas como forma de acompanhar as mudanças comportamentais de seus públicos.

Nesse contexto, a Ciência da Informação (CI), por meio de seus campos de investigação, preocupa-se com as questões relacionadas à produção, circulação e apropriação da informação, tendo em vista a criação de instrumentos e recursos, bem como o estabelecimento de metodologias e estratégias que fundamentem a mediação de informação, em suas manifestações implícita e explícita. (ALMEIDA JÚNIOR, 2015).

No cenário atual, a preocupação emergente diz respeito à obtenção de informações fidedignas de maneira minimamente organizada em curtos espaços de tempo. Dentre os estudos distintos da CI, uma área nuclear que atua diretamente nos processos de recuperação e apropriação da informação é a representação temática da informação, a qual se encontra, de modo mais amplo, no cerne da Representação da Informação. (GUIMARÃES; SALES; GRÁCIO, 2012).

A área de Representação da Informação[1] envolve distintas atividades que podem ser agrupadas em duas categorias gerais: a representação descritiva e a representação temática. As atividades que envolvem a representação dos conteúdos informacionais para fins de recuperação constituem a área de Representação Temática e se estruturaram a partir de três operações: a catalogação por assunto, a classificação e a indexação. Nessa pesquisa tomamos como foco os estudos voltados à indexação da informação.

[1] Há autores que se referem a esta área pela expressão “Tratamento da Informação”, ou ainda, de modo mais geral, “Organização da Informação”.

Segundo Redigolo e Silva (2017, p. 53) “a Indexação tem origem predominantemente inglesa” e se destacam nesses estudos “[...] os trabalhos de Foskett, Austin, Farradane, Metcalfe, Aitchinson, Gilchrist e Lancaster.” Para além destes autores, também são citados “Campos, Van Slype, Farrow, entre outros.” (SILVA; FUJITA, 2004, p. 137).

Van Slype (1991 *apud* SILVA; FUJITA, 2004, p. 137) compreende a indexação como sendo “a operação que consiste em enumerar os conceitos sobre os quais trata um documento e representa-os por meio de uma linguagem combinatória: lista de descritores livres, lista de autoridades e o thesaurus de descritores”. Essa etapa enquanto processo será responsável pela recuperação da informação, impactando diretamente na capacidade de revocação e precisão dos sistemas de informação.

Embora esta atividade seja tradicionalmente desempenhada por indexadores, no caso das bibliotecas, bibliotecários indexadores, há alguns anos diversos estudos vêm sendo desenvolvidos com o intuito de automatizar essa prática, os quais se baseiam no argumento de que a indexação manual é lenta e não dá conta da crescente produção de informação, além disso apontam que a indexação realizada pela máquina se mostra mais precisa e imparcial.

Esses argumentos são válidos, entretanto é pertinente ressaltar que a atividade de indexação é por natureza uma operação que envolve processos mentais do bibliotecário/indexador. Conforme já ressaltado por autores, como, Silva e Fujita (2004), Dias e Naves (2007), a indexação caracteriza-se como uma atividade interdisciplinar, uma vez que recebe influências de estudos linguísticos, cognitivos e lógicos (DIAS; NAVES, 2007).

Os estudos acerca da indexação manual são acompanhados de pesquisas que discutem a viabilidade e a eficácia da indexação automática e semiautomática. Tomar conhecimento do que vem sendo produzido e em qual vertente é de fundamental importância para os estudos da CI, sobretudo na área de representação temática da informação.

Considerando a longa tradição da indexação e a larga produção sobre o tema, evidenciadas em periódicos científicos, eventos e pesquisas oriundas das universidades e agências de fomento, direcionamos esta investigação ao que vem sendo produzido acerca da automatização da indexação. Para isso, tomamos como questão norteadora a seguinte indagação**:** O que há de evidências sobre a indexação automática e semiautomática, nos últimos 50 anos, na produção científica em CI?

Com base nessas reflexões iniciais e de cunho mais amplo, o objetivo geral do presente trabalho é descrever o panorama de produção científica e as tendências de pesquisa sobre os estudos voltados para a indexação automática, semiautomática e auxiliada por computador, indexada na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Brapci), no período de 1972 a 2021.

Com assente nesse objetivo geral, apoiamo-nos nos seguintes objetivos específicos: a) identificar os autores que produzem acerca dessas temáticas; b) relacionar em quais periódicos os trabalhos nessa temática vêm sendo publicados; c) cotejar as publicações sobre a temática, por ano; e d) analisar as tendências de pesquisa temática e temporalmente dentro do escopo definido.

2 REPRESENTAÇÃO TEMÁTICA DA INFORMAÇÃO

A todo momento, são elaboradas representações, as quais vão desde as formas mais rudimentares de representar às mais modernas e incrementadas pelo desenvolvimento científico e tecnológico, sobretudo das Tecnologias de Informação e Comunicação. Nesse contexto, Pinto, Meunier e Silva Neto (2008) ressaltam que:

> O significado que a palavra representação encerra não é de origem tão recente, conforme parecem imaginar alguns. Muito pelo contrário, ela sempre esteve presente no espírito humano, pelo menos, desde a Pré-história quando os homens primitivos, em suas práticas cotidianas, buscavam possibilidades de comunicação através da criação de imagens ou ideogramas; assim como da escrita cuneiforme dos sumérios e dos hieróglifos produzidos no Antigo Egito (PINTO; MEUNIER; SILVA NETO, 2008, p. 17).

Assim, é possível compreender o ato de representar como o de "indicar", o de "sinalizar algo" por meio de linguagem verbal e não-verbal. No contexto das Bibliotecas, o processo de representação se faz presente por meio da atribuição de códigos alfanuméricos, que são encontrados nas etiquetas dos livros armazenados nas estantes das bibliotecas, como forma de sinalizar, de apontar os seus documentos por meio de signos verbais e não-verbais, de acordo com suas respectivas temáticas, a fim de que sejam localizados ou recuperados em outro momento.

Conforme Saracevic (1996, p. 47), a CI, com seu enfoque contemporâneo advindo dos anos 90, preocupa-se com a questão da "[...] efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual [...]". Destarte, a representação da informação, seja em seu aspecto descritivo ou temático,

se mostra como atividade essencial para que haja a comunicação entre o indivíduo que necessita de informação e os registros do conhecimento, estejam eles em suporte físico ou digital. Isso evidencia o caráter singular que a Representação da informação exerce no contexto da Ciência da Informação e, por conseguinte, na sociedade hodierna, em que há uma crescente produção de conhecimento. Todo esse contexto só aumenta a importância das práticas de representação da informação para que os recursos informacionais se apresentem de maneira organizada por meio de esquemas sinalizadores e, consequentemente, sejam recuperados.

No contexto de instituições que atuam com a guarda, organização e disponibilização de informação como é o caso das Bibliotecas, Centros de Informação, Arquivos etc., as atividades de representação envolvem a classificação de documentos, a elaboração de resumos, a catalogação de assuntos e a indexação. Sobre esta última, objeto de nossa pesquisa, faremos uma reflexão na seção seguinte.

2.1 Indexação manual, automática, semiautomática e auxiliada por máquina[2]

Propõe-se aqui uma discussão acerca da indexação manual, prática tradicional dos bibliotecários no âmbito das bibliotecas tradicionais e digitais, em seguida é discutida a indexação automática, semiautomática e auxiliada por máquina, as quais surgem como forma de auxiliar e/ou substituir a indexação manual, mormente no contexto de rápida produção de informação digital.

2.1.1 Indexação manual

Segundo Robredo (2005, p. 165) a indexação "[...] consiste em indicar o conteúdo temático de uma unidade de informação, mediante a atribuição de um ou mais termos (ou códigos) ao documento, de forma a caracterizá-lo de forma unívoca". Embora sejam encontradas divergências no que se refere às etapas da indexação de assuntos de documentos, de modo geral, conforme define Lancaster (2004), a indexação envolve duas etapas: a "análise conceitual" e a "tradução". A primeira envolve a análise do conteúdo do documento, por meio de uma leitura especializada, com estratégias direcionadas, com vistas à identificação dos conceitos que representam a produção de conhecimento

[2] É possível encontrar na literatura da área as expressões "indexação auxiliada por máquina" e "indexação assistida por máquina" como termos sinônimos.

analisada, e a segunda etapa consiste na tradução do assunto do documento, representado por conceitos selecionados, para uma linguagem de indexação, ou seja, uma linguagem que pode envolver termos que constam no próprio documento ou termos retirados de outras fontes, caracterizados como indexação por extração ou indexação por atribuição, respectivamente.

Para a Associação Brasileira de Normas Técnicas (1992), a indexação é executada por meio de três etapas, a saber: exame do documento e estabelecimento do assunto de seu documento; identificação dos conceitos presentes no assunto e tradução desses conceitos nos termos de uma linguagem de indexação.

A indexação, mormente a "análise de assunto", caracteriza-se como uma operação eminentemente subjetiva, uma vez que envolve o ser humano por meio de processos mentais, linguísticos e lógicos em uma atividade interativa que exige, conforme ressalta Fujita (2003, p. 69), os seguintes elementos: "leitor, texto e contexto". Isto posto, verifica-se o quão complexo é o ato de indexar, uma vez que requer tempo, dedicação e capacitação. Dada a sua complexidade, surgiram estudos voltados para a automação dessa atividade, sobretudo, com o intuito de dar conta do volume de informação que vem sendo produzido em ambiente digital. Discutem-se mais adiante alguns desses estudos nas subseções seguintes.

2.1.2 Indexação automática

Antes de adentrar na prática da indexação auxiliada e executada completamente por computadores dotados de programas específicos, faz-se necessário que se exponha aqui a variedade terminológica utilizada nessa área, sendo fácil encontrar na literatura as expressões: indexação automática, indexação automatizada, indexação auxiliada por computador, indexação assistida por máquina e indexação mecânica.

Vieira (1988) se refere à indexação automática como a operação que, por meio de programas de computador, identifica as palavras ou expressões representativas dos conteúdos dos documentos. Silva e Fujita (2004, p. 145) ressaltam que a "indexação automatizada seria, portanto, aquela resultante do trabalho intelectual de um profissional para checagem do valor dos termos atribuídos a um documento por um programa de computador". Sob esta definição, a indexação automatizada refere-se à indexação realizada pela máquina com validação por um profissional. No entanto, há autores que utilizam a expressão "indexação semiautomática" para se referir a esta atividade, estar-se-

ia diante, então, na percepção das autoras supracitadas, de termos sinônimos: indexação automatizada e indexação semiautomática.

Em termos de definição pelo Dicionário Aurélio, o vocábulo “automatizado” remete a algo “que ficou automático; que passou a funcionar por um sistema de mecanização”. Corroborando essa definição, Cunha e Cavalcanti (2008, p. 39), por meio do Dicionário de Biblioteconomia e Arquivologia (2008), traz o seguinte verbete para o termo “automatização”: “Introdução, numa máquina, de um método ou sistema que lhe permita ser auto controlável e auto comandável sem a intervenção humana.

Logo, é possível compreender os termos “automático” e “automatizado” como sinônimos, sendo mais viável, no tocante ao aspecto semântico, utilizar a expressão “semiautomática” para a indexação realizada pela máquina com a validação final de um profissional.

No que se refere à indexação realizada por programas computacionais, Chaumier (1988, *apud* FREITAS JUNIOR *et al.*, 2016, p. 2) conceitua a indexação como:

> Operação que descreve e caracteriza um documento, com o auxílio da representação dos conceitos nela contidos. Tal ação pode ser feita por um indexador humano, sendo nesse caso denominada na literatura indexação manual; por um programa de computador, sendo denominada indexação automática; ou ainda por um programa de computador e depois revista por um indexador humano, sendo denominada indexação semiautomática (CHAUMIER, 1988, *apud* FREITAS JUNIOR *et al.*, 2016, p. 2).

Nesse contexto, Guimarães (2000) expõe a indexação associada ao computador em três concepções: a indexação manual apoiada por programas informáticos para que auxiliem o profissional no armazenamento dos termos de indexação identificados pelo homem; a indexação realizada pela máquina, mas com a validação dos termos por um profissional (indexação semiautomática) e a indexação realizada completamente pela máquina, sem a interferência do profissional, esta última sendo chamada de indexação automática.

As justificativas para o desenvolvimento de estudos de indexação automática se baseiam em diferentes argumentos, Borges (2009, p. 15) ressalta que “seu surgimento se deu devido à necessidade de serem resolvidos problemas como a morosidade trazida pela indexação manual. Por isso, a indexação automática é vista como uma alternativa para agilizar esse processo, através dos recursos oferecidos pela tecnologia.”. Além, disso, outro aspecto motivador para as propostas de indexação automática diz respeito à subjetividade inerente ao indexador humano. Para Gil Leiva (1999, *apud* NARUKAWA, 2011, p. 47), “A

subjetividade do indexador, aliada ao tempo gasto e ao custo alto são argumentos dos defensores da indexação automática."

De modo geral, os estudos acerca da indexação automática se iniciaram baseando-se em critérios de ocorrência/frequência de palavras, sendo as recorrentes interpretadas como significativas, eliminando as palavras consideradas vazias, como, por exemplo, os artigos, as preposições, as conjunções etc. Contudo, os pesquisadores dessa área começaram a identificar inconvenientes ligados aos métodos estatísticos e probabilísticos presentes na indexação automática. Destarte, começaram a surgir pesquisas baseadas em métodos linguísticos de indexação automática, dando ênfase à semântica das palavras e das estruturas linguísticas, como, por exemplo, os Sintagmas Nominais.

A indexação automática, semelhante ao que acontece com a indexação manual, pode ser executada de duas formas: a indexação automática por extração e a indexação automática por atribuição. Aquela é realizada com base na ocorrência e frequência de palavras presentes no documento para representar o conteúdo desse texto. Esta é realizada atribuindo termos de um vocabulário controlado, por meio do qual se alcança mais uniformidade e padronização na linguagem, além de se evitarem problemas provenientes de fenômenos da língua natural (LANCASTER, 2004).

2.1.3 Indexação semiautomática

A indexação semiautomática é realizada por programas computacionais e finalizada pelo indexador humano. Em revisão de literatura sobre a automatização no processo de indexação, conforme Gil Leiva (1999, p. 57; 2008, p. 320 *apud* NARUKAWA, GIL LEIVA, FUJITA, 2009, p. 101), foram identificadas várias expressões acerca da associação da indexação com a automatização, as quais evidenciaram a existência de três conceitos: indexação automática, assistida por computador e a semiautomática. Para os referidos autores, nesta última "Os programas realizam a análise dos documentos de modo automático e se necessário os termos são validados por um profissional".

Enquanto Pinto (2000) utiliza os termos "indexação semiautomática" e "indexação auxiliada por computador" como sinônimos, Gil Leiva diferencia, ressaltando que a "indexação auxiliada pelo computador" é aquela em que se utiliza o computador para gravar os dados (os termos de indexação), já a indexação semiautomática envolve uma atividade mais complexa pelo computador, uma vez que *softwares* realizam análises linguísticas e estatísticas nos textos, selecionando os possíveis termos representativos dos documentos,

os quais são validados pelo indexador humano. Nesta indexação há a junção do processamento rápido e a baixo custo de um grande volume de informação, sem deixar de lado o profissional indexador, o qual se torna fundamental nesse processo por motivos já expostos anteriormente. Assim, estudos que conjuguem esses dois elementos se tornam promissores, ao passo que se beneficiam das tecnologias sem se isentarem de um elemento fundamental para a essência desta atividade: o profissional indexador.

2.1.4 Indexação auxiliada por máquina

Conforme evidenciado nos parágrafos anteriores, alguns autores usam a expressão indexação semiautomática como sinônima para indexação assistida pelo computador como pode ser observado em Pinto (2000). No tocante ao uso dos termos como sinônimos, sobretudo em relação aos termos indexação automática, semiautomática e auxiliada por máquina, Borges (2009, p. 31) menciona que a indexação automática "também chamada de indexação assistida por computador e de indexação semiautomática" é compreendida como um modelo de extração baseado em métodos estatísticos e probabilísticos. Isto posto, no escopo deste trabalho, adota-se a percepção de Gil Leiva (1997), diferenciando, assim, em aspectos conceituais, a indexação semiautomática da auxiliada por máquina, sendo esta última aquela em que se utiliza o computador para gravar os dados (os termos de indexação).

3 PROCEDIMENTOS METODOLÓGICOS

Conforme Gil (2002, p. 41), as pesquisas exploratórias buscam "proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. Pode-se dizer que estas pesquisas têm como objetivo principal o aprimoramento de ideias ou a descoberta de intuições". Para Michel (2009, p. 40) a pesquisa exploratória ou bibliográfica "pode ser considerado uma forma de pesquisa, na medida em que se caracteriza pela busca, recorrendo a documentos, de uma resposta a uma dúvida, uma lacuna do conhecimento". No tocante aos meios, mediante as exposições de Gil (1993) e Michel (2009), esta pesquisa se caracteriza como exploratória e bibliográfica com fins descritivos, uma vez que descreve e detalha a produção científica sobre os temas "indexação automática", "indexação semiautomática" e "indexação auxiliada por máquina" indexada na Brapci.

A escolha pela Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (Brapci) se deu por esta ser uma base especializada na produção científica da Ciência da Informação, sendo, assim, direcionada aos propósitos deste trabalho. Diversas pesquisas no âmbito da CI vêm sendo desenvolvidas tomando como fonte esta base, além disso o domínio e a completude que a caracterizam permitem perceber que, para o eficiente desenvolvimento deste trabalho, a base em questão se mostra adequada. Segundo Bufrem, Freitas e Nascimento (2014, p. 152):

> A Brapci [...] é resultado de um projeto de pesquisa acadêmica que tem o intuito de facilitar a pesquisa de documentos e artigos da área. [...]. Constitui-se, hoje, no mais completo repositório da produção científica periódica brasileira em CI e, devido à quantidade e à confiabilidade de seus artigos, tem sido considerado ferramenta útil e segura para os pesquisadores (BRUFREM; FREITAS; NASCIMENTO, 2014, p. 152).

Ainda no tocante ao aspecto metodológico, a pesquisa foi conduzida mediante a adoção das abordagens bibliométricas. A bibliometria é o estudo dos aspectos quantitativos que envolvem a produção, disseminação, socialização e evidenciação da informação registrada (MACIAS-CHAPULA, 1998). A bibliometria vem sendo utilizada nas diversas áreas do conhecimento como metodologia para a obtenção de indicadores de avaliação da produção científica, buscando compreender o comportamento de diferentes domínios. Nessa perspectiva metodológica, a abordagem bibliométrica conduz para a identificação das características temáticas da área estudada, a saber: a prática da indexação auxiliada e executada por máquinas.

É pertinente que se ressalte, ainda, a diferentes possibilidades de procedimentos que permeiam as técnicas bibliométricas. No contexto da Ciência da Informação, esses procedimentos estatísticos se materializam por meio das técnicas bibliométricas: a Lei de Lotka permite mensurar produtividade dos cientistas; a Lei de Bradford estima o grau de relevância de periódicos em determinados domínios, ou seja, com base nessa lei, os periódicos que produzem o maior número de artigos sobre dado assunto constituem um núcleo de periódicos considerados, supostamente, como de maior qualidade ou relevância para a área temática estudada. Por fim, a Lei de Zipf permite analisar a distribuição da frequência de termos presentes em um texto. O uso dessas técnicas bibliométricas no contexto da CI permitiu a concretização dos dados quantitativos, mapeados e apresentados por meio de 5 (cinco) gráficos. Fazendo uso desses instrumentos bibliométricos, buscou-se mapear a produção sobre Indexação automática e semiautomática sob os seguintes aspectos: quantidade de artigos indexados na Brapci sobre o tema, datas de publicações;

os autores e, dentre estes, os que mais têm produzido sobre o assunto em questão e, por fim, a análise das propostas que vem sendo desenvolvidas no âmbito da automação da atividade de indexação de documentos.

A pesquisa na interface da Brapci foi realizada utilizando os seguintes termos de busca [indexação automática], [indexação mecânica], [indexação automatizada], [indexação semiautomática], [indexação semi-automática], [indexação assistida por computador] e [indexação auxiliada por computador]. Embora, conforme as regras gramaticais da Língua Portuguesa, o vocábulo "semi-automática" não seja escrito com o hífen, visto que ele não é mais utilizado em palavras formadas por prefixos (ou falsos prefixos) terminados em vogal + palavras iniciadas por outra vogal, realizamos buscas com esse termo, pois havia a possibilidade de indexação ter sido feita à época com o referido descritor. A utilização dos termos mencionados foi aplicada nos seguintes filtros: título e palavras-chave. A busca foi realizada entre o período de 1972 a 2021.

3.1 Análise e discussão

Após a identificação e análise de cada artigo científico recuperado, retiraram-se as palavras-chave que pertenciam a cada artigo com vistas à análise desses descritores. Com as expressões [indexação auxiliada por máquina], [indexação assistida por máquina] e [indexação mecânica] não foram recuperados nenhum documento. Com a expressão [indexação automatizada] recuperou-se 1 documento[3], já com a expressão [indexação automática] recuperaram-se 44 documentos**,** totalizando 45 documentos analisados. Ao utilizar a expressão [indexação semi-automática] no filtro título, não se identificou nenhum documento, já a mesma expressão no filtro palavras-chave permitiu a recuperação de 1 documento, o qual já tinha sido recuperado por meio da expressão indexação automática. Por fim, a expressão **indexação semiautomática** não permitiu a recuperação de nenhum documento pelos filtros já mencionados. Identifica-se o uso majoritário da expressão **indexação automática** na base Brapci, embora, alguns trabalhos recuperados pela referida expressão tratassem de indexação semiautomática, por exemplo.

[3] O único trabalho que utilizou a expressão "indexação automatizada" no título e nas palavras-chave foi este: NARUKAWA, C. M.; GIL-LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software sisa com uso da terminologia decs na área de odontologia. **Informação & Sociedade: Estudos**, v. 19, n. 2, 2009.

Com base nas expressões de busca mencionadas, recuperaram-se 45 documentos. Desse total, identificou-se apenas um artigo com inconsistência[4] que foi indexado na Brapci como sendo elaborado por um autor, entretanto, ao acessar o artigo na íntegra, identificamos que se tratava de dois autores. Após a separação dos autores de cada publicação encontrada, identificou-se um total de 53 autores, logo, com vistas a oferecer uma visualização melhor dos dados, adotou-se como critério elencar os autores com duas ou mais publicações, para que fosse possível a apresentação em gráfico.

Gráfico 1 – Quantitativo de publicações para cada autor

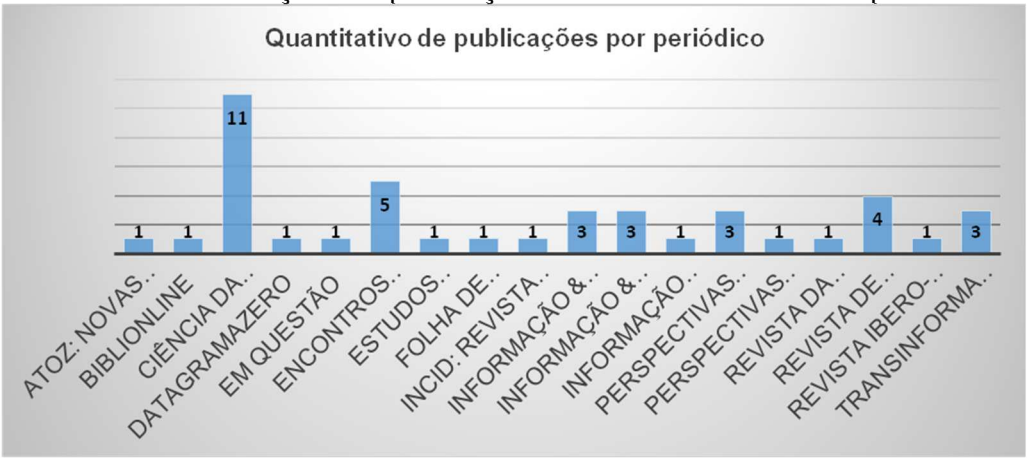
Fonte: Dados da pesquisa (2021)

Por meio do gráfico 1 é possível perceber um destaque considerável para o pesquisador Renato Fernandes Corrêa, que acumula um total de 13 publicações em português indexadas na Brapci, as quais tratam da temática geral: indexação automática, semiautomática e auxiliada por computador, embora, ao analisar os resumos das publicações do referido autor, identifica-se que a ênfase em suas pesquisas é voltada eminentemente para a indexação automática, sobretudo, por meio de Sintagmas Nominais. Na sequência, buscou-se identificar os tipos de publicações de comunicação científica presentes nas 45 publicações: Apresentação oral / comunicação oral: 2 trabalhos; Artigo / Artigo científico / Artigo de pesquisa: 34 trabalhos; Artigo de revisão: 2 trabalhos; Comunicação / Comunicação de trabalhos: 3 trabalhos; Relato de experiência: 1 Trabalho e Relato de pesquisa: 3 trabalhos.

No gráfico 2, logo abaixo, expõe-se a distribuição dos periódicos que mais publicaram sobre a temática em questão.

[4] Estudo comparativo de indexação de texto completo para recuperação de informações em sistemas gerenciadores de banco de dados. Autores: Edson Marchetti da Silva e Lucas Meneses Mardegan. Disponível em: https://brapci.inf.br/index.php/res/v/114978.

Gráfico 2 – Distribuição das publicações analisadas conforme os periódicos



Fonte: Dados da pesquisa (2021)

Com base no gráfico 2, as revistas que mais publicaram sobre a temática foram, respectivamente: Ciência da Informação (11 publicações), Encontros Bibli: revista eletrônica de Biblioteconomia e Ciência da Informação (5 publicações) e a Revista de Biblioteconomia de Brasília (4 publicações). O destaque foi da revista Ciência da Informação, responsável pelo maior número de publicações sobre essa temática, esse dado pode estar relacionado com o tempo de existência dessa revista, da qual o primeiro fascículo foi lançado em 1972. Essa década foi marcada pela criação dos programas de pós-graduações brasileiros, inclusive o de Ciência da Informação, vinculado ao Instituto Brasileiro de Informação em Ciência e Tecnologia. Pelo seu longo tempo de existência podemos inferir uma relação direta com a quantidade de publicações, tendo em vista que esta é “considerada a revista mais tradicional e importante do país sobre a área da CI” (SANTOS; COSTA, 2012, p. 105).

Os outros dois periódicos com mais publicações foram a Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação e a Revista de Biblioteconomia de Brasília, as quais tiveram seus inícios, respectivamente, em 1973 e 1996. Dos 45 documentos recuperados na Brapci, dois foram veiculados em eventos, um no XVIII Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB, e o outro no Encontro Brasileiro de Bibliometria e Cientometria. Na sequência das análises, são expostos os anos das publicações coletadas, com vistas a identificar os períodos mais produtivos para a temática em estudo.

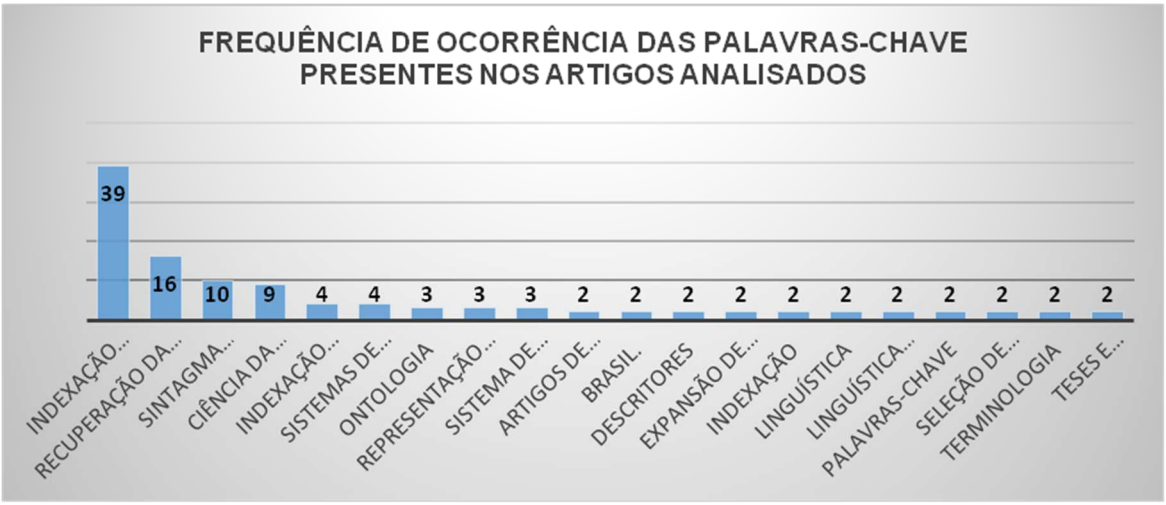
Gráfico 3 – Distribuição das publicações entre os anos de 1980 e 2020

Fonte: Dados da pesquisa (2021)

O primeiro artigo registrado sobre a temática foi publicado no ano de 1980. Identificamos, com base no gráfico acima, um aumento nas pesquisas sobre indexação automática, semiautomática e indexação auxiliada por computador nos anos de 1988 e 1991, tendo um leve aumento apenas em 2009, evidenciando um espaço de tempo de 12 anos com uma média de apenas uma publicação por ano. Percebeu-se a partir de 2013, sobretudo nos últimos anos (2017, 2018, 2019 e 2020), uma crescente na produção de pesquisas voltadas a esta temática e inferimos ser este um período de evidência para os estudos em Indexação automática por atribuição, fazendo uso de vocabulários controlados e baseando-se em sintagmas nominais como fontes de informação, bem como também propostas de indexação social, folksonomia assistida, além de estudos associando a indexação automática à ontologia.

Por fim, analisaram-se, com base nas palavras-chave presentes em cada publicação, as temáticas mais frequentes, por meio dos termos mais recorrentes utilizados para sinalizar os conteúdos dos documentos, os quais, por sua vez, representam os conteúdos das publicações, conforme mostra o gráfico 4.

Gráfico 4 - Termos mais recorrentes utilizados como palavras-chave nos documentos analisados.



Fonte: Dados da pesquisa (2021)

Para elaboração do gráfico 4, utilizou-se um ponto de corte de no mínimo 2 ocorrências para que o descritor figurasse no referido gráfico. Foi possível identificar que dentre os termos analisados o mais recorrente foi de fato “indexação automática” (com 39 aparições), além disso percebeu-se, por meio da leitura dos resumos dos referidos trabalhos, que os estudos voltados para essa temática estão diretamente ligados a metodologias, softwares e estratégias de indexação automática, ou seja, metodologias de indexação totalmente realizada pela máquina, sem intervenção do indexador humano.

Assim, por meio da análise dos descritores utilizados na indexação do corpus selecionado foi possível perceber os termos de indexação mais recorrentes, demonstrando, assim, as temáticas mais exploradas pelas pesquisas. Considerando os termos com uma frequência de ocorrência acima de 3 vezes, temos os seguintes descritores: Indexação automática; Recuperação da informação; Sintagmas nominais; Ciência da Informação; Indexação automática por atribuição; Sistemas de recuperação da informação; Ontologia e Representação da informação.

As pesquisas de Kuramoto (1996 e 1999) se apresentam como marcos no estudo da viabilidade de uso dos Sintagmas Nominais em sistemas de Recuperação da Informação. A partir desses estudos, outras pesquisas vêm se debruçando sobre as potencialidades dos SNs para a indexação automática de documentos. Diversos autores, como, por exemplo, Le Guern (1991), Kuramoto (1995; 1996; 2002), Souza (2005: 2006), Maia (2008), Morellato (2010), Corrêa et al. (2011), Silva e Correa (2015), Nascimento (2015), Corrêa e Bazílio (2017), Nascimento e Corrêa (2018; 2019), Corrêa e Celerino (2019) desenvolveram pesquisas voltadas à indexação baseada em Sintagmas Nominais.

Na sequência, o gráfico 5 nos mostra grupos de termos, dentre os quais estão os três utilizados na pesquisa inicial deste trabalho, a saber: Indexação automática, Indexação semiautomática e Indexação auxiliada/assistida por máquina. **O grupo 1** engloba os seguintes termos: Indexação Automática / Sistema de indexação automática / Avaliação da indexação automática / Avaliação de sistemas de indexação automática / Indexação automática por atribuição / Indexação automática e manual / indexação automatizada, Software de indexação automática, Indexação automática derivativa e Critérios de indexação automática. **O grupo 2** engloba os seguintes termos: Sintagma Nominais / Seleção de Sintagmas nominais. **O grupo 3** engloba os seguintes termos: Indexação Semi-automática /Sistema de indexação semi-automático. **O grupo 4** envolve os seguintes termos: Indexação Auxiliada por máquina / Indexação assistida por máquina.

Gráfico 5 – Frequência de ocorrência de termos específicos

Fonte: Dados da pesquisa (2021)

O gráfico acima, por meio da frequência de termos indexados na base analisada, mostra uma crescente nos estudos voltados para a indexação automática com uma inclinação para a indexação automática baseada em Sintagmas Nominais. Comparando esta indexação com as propostas de indexação semiautomática, verifica-se uma distância considerável, visto que apenas duas ocorrências de termos voltados para propostas semiautomáticas foram identificadas.

4 CONCLUSÃO

Os estudos voltados à produção científica se mostram singulares, uma vez que possuem o potencial de sinalizar as temáticas mais pesquisadas dentro de determinada área, os principais autores e as tendências de pesquisa, as quais nos mostram projeções para futuras investigações. Retomando os objetivos a que se propôs esta pesquisa: “a) identificar os autores que produzem acerca dessas temáticas; b) relacionar em quais periódicos os trabalhos nessa temática vêm sendo publicados; c) cotejar as publicações

sobre a temática, por ano; e d) analisar as tendências de pesquisa temática e temporalmente dentro do escopo definido", considera-se que este estudo os alcançou uma vez que descreveu o campo de estudo investigado com base na Brapci. Verificou-se que a produção científica sobre a temática em questão é proveniente de estudos realizados por distintos autores, com ênfase para o pesquisador Renato Fernandes Corrêa, o qual apresenta destaque nessa vertente.

No que se refere ao periódico com mais publicações, ganha destaque a revista Ciência da Informação, esse comportamento pode estar relacionado ao tempo de existência dessa revista, da qual o primeiro fascículo foi lançado em 1972. Em relação aos períodos mais férteis para o domínio analisado, verificou-se um aumento nas pesquisas sobre indexação automática, semiautomática e indexação auxiliada por computador na década de 90, ficando um espaço de tempo com poucas produções, retomando a partir de 2013, sobretudo nos últimos anos (2017, 2018, 2019 e 2020), uma crescente na produção de pesquisas voltadas a esta temática.

Além dos dados quantitativos que descrevem o comportamento do domínio ao longo do tempo, buscou-se analisar as tendências de pesquisas voltadas à automação da indexação. Percebe-se que as pesquisas nessa área se dedicam, em sua maioria, a propostas e estratégias de capacitação da máquina para a realização da indexação de forma independente do profissional indexador, ou seja, a busca pela automatização de forma integral da indexação.

Esses estudos ganharam melhores performances a partir do momento em que se utilizaram não mais as palavras soltas de um texto, mas as menores unidades portadoras de significado específico, ou seja, os sintagmas nominais, evidenciando, desta forma, a utilização da semântica embutida nos próprios documentos, embora haja avanços nesta área de estudo, muito ainda há de se pesquisar para que uma máquina, por meio de softwares projetados para determinadas atividades de representação da informação, consiga desempenhar as operações tal qual o indexador humano. Estudos voltados para essa área apoiam-se nas pesquisas de Inteligência Artificial – IA, as quais buscam capacitar as máquinas para se comportarem de forma semelhante ao ser humano diante de operações e atividades específicas.

A natureza subjetiva da indexação é evidente conforme já ressaltado por Fujita (2003, p. 69), para a qual os seguintes elementos: "leitor, texto e contexto" caracterizam essa subjetividade, o que noz faz perceber o desafio que é capacitar a máquina para agir como o indexador humano, por outro lado são inquestionáveis os benefícios trazidos pelas

tecnologias, embora se reconheçam suas limitações. Isto posto, evidencia-se que as pesquisas, que conjugam essas duas vertentes cotejando as possibilidades das tecnologias com a capacidade ímpar do indexador humano, se mostram frutíferas para a área da representação, visto que se beneficiam do tratamento automático de grandes volumes de informações, selecionando possíveis descritores documentais, sem com isso deixarem de ter a validação final por um indexador humano.

Benefícios advindos da Inteligência Artificial - IA contribuem incisivamente para o processamento de grandes volumes de informação no meio digital, uma vez que se alcança um processo uniforme de diferentes documentos em curtos espaços de tempo, nessa vertente, destaca-se o *Machine Learning (ML)*, subcategoria da IA que vem se mostrando promissora como recurso que permite aos computadores desenvolverem o reconhecimento de padrões, aprendendo de forma contínua e fazendo previsões com base em dados e até fazendo ajustes em seu comportamento, dispensando uma nova programação para isso.

Há, ainda de forma incipiente, estudos voltados para a inteligência artificial aplicada à indexação automática de imagens, apoiando-se em estratégias de *Deep Learning.* Assim, a *Machine Learning (ML)* pode contribuir incisivamente à indexação automática, uma vez que possibilitará a esta mais autonomia no processamento da informação, evidenciando propostas promissoras na prática da representação temática da informação de forma automática em curtos espaços de tempo.

REFERÊNCIAS

ALMEIDA JÚNIOR, O. F. de. Mediação da informação: um conceito atualizado. In: BORTOLIN, S.; SANTOS NETO, J. A.; SILVA, R. J. (Org.). **Mediação oral da Informação e da leitura**. Londrina: Abecin, 2015. p. 9-32.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12.676**: Métodos para análise de documentos – determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro: ABNT, 1992.

BORGES, G. S. B. **Indexação automática de documentos textuais**: proposta de critérios essenciais. 2009. 113 f. Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Minas Gerais, 2009.

BUFREM, L. S.; FREITAS, J. L.; NASCIMENTO, B. S. Autoria e pesquisa em organização do conhecimento: análise da produção científica em ciência da informação. **Em Questão**, Porto Alegre, v. 20, n. 3, 2014. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/49281. Acesso em: 29 ago. 2021.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008. xvi, 451 p.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116 p. (Estudos Avançados em Ciência da Informação, 3).

FERREIRA, Aurélio Buarque de Holanda. **Novo Aurélio século XXI**: o dicionário da língua portuguesa. 3 Curitiba: Editora Positivo, 2004, 2120 p.

FREITAS JÚNIOR, N. et al. Indexação semiautomática de publicações através de técnicas de mineração de texto. In: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 12, 2016, Rio de Janeiro. **Anais**... Rio de Janeiro: UFF, 2016. Disponível em: http://www.inovarse.org/sites/default/files/T16_222.pdf. Acesso em: 04 set. 2021.

FUJITA, M. S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia & Ciência da Informação**, v. 1, n. 1, p. 60-90, 2003. Disponível em: https://brapci.inf.br/index.php/res/v/40149. Acesso em: 15 set. 2021.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas S/A, 2002.

GUIMARÃES, J. A. C. **Indexação em um contexto de novas tecnologias**. [S.l.: s.n.], 2000. 10 p. Texto Didático.

GUIMARÃES, J. A. C.; SALES, R.; GRÁCIO, M. C. C. A dimensão interdisciplinar da análise documental nos contextos brasileiro e espanhol no âmbito da organização do conhecimento. **DataGramaZero**, v. 13, n. 6, 2012. Disponível em: http://hdl.handle.net/20.500.11959/brapci/7992. Acesso em: 15 set. 2021.

LANCASTER, F. W. **Indexação e Resumos**: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. 2. ed. revista e atualizada. Brasília, DF: Briquet de Lemos, 2004.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, v. 27, n. 2, p. 134-140, 1998.

MICHEL, M. H. **Metodologia e pesquisa científica em Ciências Sociais**: um guia prático para acompanhamento da disciplina e elaboração de trabalhos monográficos. 2. ed. São Paulo, Atlas, 2009.

NARUKAWA, C. M. **Estudo de vocabulário controlado na indexação automática**: aplicação no processo de indexação do Sistema de Indización Semiautomática (SISA). 2011. 222 f. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2011. Disponível em: https://repositorio.unesp.br/handle/11449/93677. Acesso em: 04 set. 2021.

NARUKAWA, C. M.; GIL-LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software sisa com uso da terminologia decs na área de odontologia. **Informação & Sociedade:** Estudos, v. 19, n. 2, 2009. Disponível em: http://hdl.handle.net/20.500.11959/brapci/91939. Acesso em: 15 set. 2021.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado. **Perspectivas em Ciência da Informação**, v. 6, n. 2, 2000. Disponível em: http://hdl.handle.net/20.500.11959/brapci/37708. Acesso em: 04 set. 2021.

PINTO, V. B.; MEUNIER, Jean-Guy; SILVA NETO, C. A contribuição peirciana para a representação indexal de imagens visuais. **Enc. Bibli**: R. Eletr. Bibliotecon. Ci. Inf., Florianópolis, n. 25, p. 15-35, 1º sem. 2008. Disponível em: http://www.periodicos.ufsc.br/index.php/eb/article/view/1153/878 Acesso em: 31 ago. 2021.

REDIGOLO, F. M.; SILVA, M. V. A representação temática como mediadora implícita da informação em bibliotecas universitárias. **Ponto de Acesso**, v. 11, n. 2, p. 49-69, 2017. DOI: 10.9771/rpa.v11i2.14307. Acesso em: 29 maio 2022.

ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. e ampl. Brasília: Edição de autor, 2005.

SANTOS, R. M. G. D.; COSTA, L. F. Usabilidade na ciência da informação: uma análise da produção científica. **Prisma.com (Portugal)**, n. 19, p. 97-124, 2012. Disponível em: http://hdl.handle.net/20.500.11959/brapci/64674. Acesso em: 11 set. 2021.

SARACEVIC, T. Ciência da informação: origem, evolução e relações **Perspec. Ci. Inf**., Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: https://www.brapci.inf.br/_repositorio/2017/07/pdf_7810a51cca_0000015436.pdf. Acesso em: 09 ago. 2021.

SILVA, M. R.; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas. v. 16, n. 2, p. 133-161, maio./ago., 2004. Disponível em: https://www.scielo.br/j/tinf/a/cNngvqQdWfBGrJtLSdLRKnP/?format=pdf&lang=pt. Acesso em: 13 jul. 2021.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, v. 17, n. 1, 1988. Disponível em: https://brapci.inf.br/index.php/res/v/20414. Acesso em: 28 ago. 2021.

NOTAS

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: G. D. Nascimento, G. K. M. Gonçalves, M. E. B. C. Albuquerque
Coleta de dados: G. D. Nascimento, G. K. M. Gonçalves, M. E. B. C. Albuquerque
Análise de dados: G. D. Nascimento, G. K. M. Gonçalves, M. E. B. C. Albuquerque
Discussão dos resultados: G. D. Nascimento, G. K. M. Gonçalves, M. E. B. C. Albuquerque
Revisão e aprovação: G. D. Nascimento, G. K. M. Gonçalves, M. E. B. C. Albuquerque

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Mayara Madeira Trevisol, Jônatas Edison da Silva, Camila Letícia Melo Furtado e Beatriz Tarré Alonso.

HISTÓRICO
Recebido em: 28-11-2022 – Aprovado em: 20-02-2023 – Publicado em: 10-04-2023