ARTIGO

Recebido em:
26/09/2017

Aceito em:
18/06/2018

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 23, n. 53, p. 64-77,set./dez., 2018. ISSN 1518-2924. DOI: 10.5007/1518-2924.2018v23n53p64

A consistência na indexação automática por atribuição de artigos científicos na área de Ciência da Informação

The consistency in the automatic indexing by assignment of scientific articles in the area of Information Science

Marcio Aercio Silva BANDIM (aercios49@gmail.com) *
Renato Fernandes CORRÊA (renato.correa@ufpe.br) **
* Mestre em Ciência da Informação pela Universidade Federal de Pernambuco.
** Professor Doutor da Universidade Federal de Pernambuco, no Departamento de Ciência da Informação e no Programa de Pós-graduação em Ciência da Informação – PPGCI–UFPE.

Resumo
Avalia a qualidade da indexação automática por atribuição de artigos científicos brasileiros na área de Ciência da Informação via índice de consistência com a indexação intelectual. Trata-se de uma pesquisa exploratória e experimental. O experimento computacional constitui da indexação automática por atribuição do corpus de 60 artigos científicos selecionados por Souza (2005), aplicando o Sistema de Indización Semi-Automático (SISA) com os descritores do Tesauro Brasileiro de Ciência da Informação (TBCI). Os termos atribuídos pelo SISA foram comparados com as palavras-chave dos autores, adotando-se o critério de consistência relaxada. Foi obtido um índice de consistência médio de 14% com a atribuição de termos gerais habilitada na configuração do SISA, e 19% sem habilitar a atribuição de termos gerais. Conclui-se que a indexação automática por atribuição de artigos do corpus tem desempenho satisfatório em termos de índice de consistência.
Palavras-chave: Indexação automática por atribuição. Palavras-chave. Tesauro Brasileiro de Ciência da Informação. Consistência na indexação. Artigos de periódico científico.

Abstract
This work evaluates the quality of automatic indexation by attribution of articles published in Brazilian scientific journals of the Information Science area through consistency indices with the manual indexing. It is an exploratory and empirical research. The computational experiment uses the corpus of 60 scientific articles selected by Souza (2005), the descriptors of Brazilian Thesaurus of Information Science (TBCI), and the proposed terms by the Semi-Automatic Indexing System (SISA) for automatic indexing by assignment of the documents. A comparison between SISA terms and authors' keywords results in an average of 14% for the consistency index, using the relaxed consistency criterion and the General Terms file enabled in the SISA configuration. Without enabling this file, results an average of 19% for the consistency. Thus, we could conclude that automatic indexing by assignment of articles of the corpus has satisfactory performance by consistency index.
Keywords: Automatic indexing by assignment. Keywords. Brazilian Thesaurus of Information Science. Indexing consistency. Scientific journal articles.

v. 23, n. 53, 2018.
p. 64-77
ISSN 1518-2924

1 INTRODUÇÃO

A organização e recuperação da informação são disciplinas importantes no estudo e pesquisa sobre instrumentos para divulgação e acesso ao conhecimento científico. Entre os instrumentos que potencializam a comunicação científica, se destacam as bases de dados de publicações científicas, que realizam os processos de indexação do conteúdo ou assunto destas publicações, objetivando disponibilizar o acesso via recuperação da informação.

O processo de indexação faz parte dos procedimentos que compõem a Análise Documentária, que pode ser definida como "[...] a operação pela qual escolhe-se os termos mais apropriados para descrever o conteúdo de um documento. Este conteúdo é expresso pelo vocabulário da linguagem documental escolhida pelo sistema e os termos são ordenados para construir índices que servirão à pesquisa". (GUINCHAT; MENOU, 1994, p. 175).

Quando se trata da organização e recuperação da informação, a análise automática de textos é estudada e pesquisada desde a década de 1950. Segundo Robredo (2005), para disponibilizar o acesso rápido às bases de dados científicas, o suporte do computador é de suma importância no processamento de dados e informações.

Assim surgiu a indexação automática, definida por Lapa e Corrêa (2014), como um conjunto de operações matemáticas, linguísticas e de programação que, quando aplicada na análise de documentos, faz o processamento dos conteúdos, selecionando automaticamente os termos representativos dos assuntos destes documentos, termos estes que serão usados na recuperação da informação.

Em termos de indexação com o uso do computador, existe a indexação automática por extração, a indexação automática por atribuição, e a indexação semiautomática.

Neste trabalho, é discutida a indexação automática por atribuição que, segundo Lancaster (2004), consiste numa representação temática por meio de termos selecionados de um vocabulário controlado (tesauro ou lista alfabética), onde um programa de computador desenvolve para cada termo a ser indexado um "perfil" de palavras ou expressões.

Pesquisa-se a indexação automática por atribuição por esta permitir o controle terminológico através do uso de vocabulário controlado.

No processo de indexação existem métricas que permitem avaliar a qualidade da indexação. Entre estes critérios pode ser citada a consistência na indexação.

Segundo Leiva (2001, p. 81) a consistência na indexação é definida como o grau de concordância na representação da informação essencial de um documento, através de um conjunto de termos de indexação, selecionados por cada um dos indexadores de um grupo ou por um sistema de indexação.

Embora a consistência não seja o único aspecto determinante da qualidade da indexação (LANCASTER, 2004), no contexto deste trabalho, tendo as palavras-chave dos autores como termos de indexação a serem comparados com termos de indexação atribuídos por um sistema de indexação, o índice de consistência se configura como um bom indicador da qualidade da indexação automática.

Assim, o problema de pesquisa deste trabalho consiste na necessidade de se investigar o processo e a qualidade resultante na representação do assunto mediante a indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação.

Neste contexto, a problemática desta pesquisa pode ser resumida assim: A indexação automática por atribuição de artigos de periódicos científicos na área de Ciência da Informação resulta em valores satisfatórios para o índice de consistência? Os termos das palavras-chave definidas nos artigos científicos pelos autores contribuem para obtenção de valores satisfatórios para o índice de consistência, quando comparados com os termos de indexação propostos pela indexação automática por atribuição?

Deste modo, tem-se como objetivo geral neste trabalho: avaliar a qualidade da indexação automática por atribuição de artigos científicos brasileiros na área de Ciência da Informação via índice de consistência. Para tanto, foram definidos como objetivos específicos:

a) avaliar a aplicação da indexação automática por atribuição, usando o Sistema de Indización Semi-Automático (SISA) como sistema de indexação

automática e o Tesauro Brasileiro da Ciência da Informação (TBCI) como vocabulário controlado;

b) comparar os termos de indexação propostos pelo SISA com os termos das palavras-chave dos 60 artigos científicos referenciados na tese de Souza (2005) e,

c) analisar a qualidade da indexação automática por meio do índice de consistência.

No contexto deste trabalho, o índice de consistência retrata o percentual de concordância entre os termos de indexação propostos pelo SISA via indexação automática por atribuição e os termos da indexação manualmente atribuídos pelos autores via palavras-chave.

A justificativa para realização deste trabalho está na importância da avaliação da qualidade da indexação automática por atribuição de artigos de periódicos da área de Ciência da Informação. Isto permite, analisar, sua aplicabilidade no contexto da criação de bases de dados científicas na área, bem como, desperta reflexões quanto à metodologia para avaliação da qualidade da indexação automática e possíveis aperfeiçoamentos dos métodos de indexação automática por atribuição.

2 REFERENCIAL TEÓRICO

Para um panorama mais amplo das pesquisas sobre indexação automática no Brasil, recomenda-se a leitura do trabalho (CORRÊA; LAPA, 2013).

Embora o assunto principal do presente trabalho seja a indexação automática por atribuição de artigos científicos em língua portuguesa, descreve-se inicialmente trabalhos correlacionados que abordam o uso de linguagem controlada na indexação de publicações em bases de dados científicas, como por exemplo o de Fujita (2004).

Fujita (2004) afirma que a palavra-chave é uma representação do conteúdo significativo do texto, sendo também utilizada para representar uma necessidade de informação na estratégia de busca. A palavra-chave está sempre associada à busca de informação em bases de dados, por serem representativas do conteúdo dos documentos armazenados.

Citando Kobashi (1994), a autora supracitada lembra que a descrição de conteúdo do documento é denominada de Análise Documentária a qual é composta de três etapas: Análise, Síntese e Representação. É interessante observar que a Representação é construída por meio de um processo de condensação intensiva do texto original gerando os diferentes tipos de resumo e a determinação de palavras-chave como produtos documentários.

Segundo Fujita (2004) o texto pode ser representado pelo resumo e, no menor nível de condensação, pelas palavras-chave ou número de classificação. Assim, o resultado da indexação, da classificação e da elaboração de resumos se constitui numa representação do conteúdo do documento para sua acessibilidade temática.

Fujita (2004) enfatiza a importância da indexação na análise documentária lembrando que, sob o ponto de vista dos sistemas de recuperação, essa condiciona os resultados de uma estratégia de busca. As palavras-chave ou descritores podem ser determinados fazendo-se uso de um código comutador, ou seja, uma linguagem documentária que poderá ser um tesauro ou lista de cabeçalho de assunto especializados na área do assunto objeto da indexação.

Assim, o processo de indexação que resultará na determinação de palavras-chave, está vinculado a uma metodologia de análise de assunto, que combina a exploração da estrutura textual do artigo com uma sistemática de identificação de conceitos. Aliada a isto, tem-se, também, a representação documentária, realizada mediante compatibilização com linguagens documentárias dos termos identificados e selecionados durante a análise de assunto.

Outro trabalho correlacionado é de autoria de Miguéis e colegas (MIGUÉIS et al., 2013), que tem como objetivo analisar as palavras-chave dos autores nos artigos científicos, comparando-as com os termos do tesauro Medical Subject Headings (MeSH) atribuídos aos mesmos na base de dados MEDLINE. Os autores destacam a importância das palavras-chave por exporem a abrangência de um assunto e seus conceitos principais, bem como serem úteis para indexação em mecanismos de busca ou categorização de textos.

As pesquisas sobre a importância e caraterísticas das palavras-chave tem abrangido vários aspectos (MIGUÉIS et al., 2013), como: o da eficiência na recuperação da informação; o uso para a extração automática de termos a partir de diferentes metodologias e algoritmos; o uso por parte dos autores e editores; a sua utilização nas etiquetagens (*metatags*); e a comparação com os títulos, resumos e textos integrais.

Os autores supracitados abordam a comparação das palavras-chave descritas pelos autores de artigos científicos com os descritores empregados por indexadores, visando analisar a contribuição que estas podem dar para a escolha dos termos de indexação. Para isso, os autores delimitaram uma amostra, constituída por 182 artigos científicos publicados entre 1996 e 2012 em revistas internacionais da área das Ciências Farmacêuticas, depositados no repositório Estudo Geral da biblioteca da Faculdade de Farmácia da Universidade de Coimbra.

Assim, a escolha de uma linguagem controlada recaiu no tesauro Medical Subject Headings (MeSH) da MEDLINE – principal base de dados bibliográficos de artigos científicos da área das Ciências da Saúde, com cerca de 19 milhões de referências, publicadas desde 1948.

Na comparação dos termos das palavras-chave com os descritores do tesauro MeSH, os autores fazem uma análise quantitativa. Assim, determinaram que as relações consideradas de equivalência, corresponderiam à classificação de 1 (um) ponto e, as relações de semelhança seriam classificadas com meio ponto (0,5).

No reconhecimento das palavras-chave equivalentes aos descritores do tesauro foram definidos os seguintes critérios: termos simples com a mesma grafia; termos com variantes em gênero ou número, termos compostos com a mesma grafia, ainda que com ordem invertida, e termos com o mesmo significado do ponto de vista semântico. Para as outras palavras-chave nas quais se identificou algum tipo de relação semântica com os termos escolhidos pelo MeSH, estas foram consideradas como termos associados.

Após a análise e avaliação da comparação dos termos das palavras-chave dos 182 artigos com os termos do tesauro MeSH, os autores concluíram que a maioria dos artigos registra cinco palavras-chave ou menos, e que metade das palavras-chave apresentam relações de equivalência ou associatividade com os descritores da MEDLINE, em proporções semelhantes.

A frequência das relações estabelecidas pelas palavras-chave situa-se, majoritariamente, entre 1 a 2 conceitos equivalentes e 1 a 2 conceitos associados por artigo. Nos 182 artigos analisados, 26% das palavras-chave introduzidas são equivalentes e 28% são associadas aos descritores da MEDLINE, o que representa um conjunto de 45% das palavras-chave existentes nos artigos. É preciso registrar que neste trabalho os autores não fizeram uso de software de indexação automática e compararam a indexação dos autores com a realizada pelos indexadores da MEDLINE.

Sobre a temática indexação automática por atribuição de artigos científicos em língua portuguesa, buscando-se na literatura nacional e internacional em Ciência da Informação, encontrou-se os trabalhos relacionados descritos a seguir.

Cita-se inicialmente o relato de pesquisa de Narukawa, Leiva e Fujita (2009), onde os autores analisaram a aplicação do software SISA com uso da terminologia DeCS na indexação automática de artigos de periódico na área de Odontologia.

Os autores supracitados analisaram comparativamente a indexação automática do SISA e a indexação manual do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME), mensurando a consistência na indexação, bem como a revocação e precisão na recuperação da informação.

Segundo os autores, o software SISA foi desenvolvido na Espanha por Leiva (1999), sendo inicialmente proposto para a área de Biblioteconomia e Documentação. No entanto, em virtude da flexibilidade do sistema, é possível adaptar sua configuração para ser aplicado em qualquer área e idioma para a qual se possua uma linguagem documentária e uma lista de palavras vazias no idioma do texto dos documentos. O SISA é um sistema semiautomático que analisa as partes do documento que estão delimitados com marcadores (título, resumo e texto), aplica critérios de seleção de termos e propõe os termos selecionados para indexação a serem posteriormente validados pelo indexador.

Após comparar a indexação automatizada do SISA com a indexação manual do BIREME, Narukawa, Leiva e Fujita (2009) relatam como principais conclusões:

a) A importância da estruturação adequada das fontes ou arquivos (artigo científico, lista de descritores e lista de palavras vazias) para adequado funcionamento do SISA.
b) A falta de flexibilidade na indexação automática impediu a atribuição de termos relevantes para indexação, causada principalmente pela incompatibilidade existente entre os termos do artigo científico identificados pelo software SISA e os termos da linguagem documentária DeCS. Além disso, o SISA atribuiu muitos termos simples e teve dificuldade em atribuir termos compostos.
c) Para alguns documentos não houve atribuição pelo SISA de descritores do DeCS equivalentes às palavras-chave, o que tornou inviável a recuperação de alguns artigos científicos e influenciou negativamente no valor dos índices de consistência, revocação e precisão.
d) Há necessidade de estudos em torno da adequação da linguagem documentária ao uso do software SISA, a partir da incorporação e avaliação de métodos linguísticos de análise morfológica e sintática. Para isso, há necessidade da convergência do conhecimento de diversas áreas, entre as quais, a Linguística computacional, a Ciência da Computação, a Estatística e a Ciência da Informação, para o pleno desenvolvimento da área de automatização da indexação.

O trabalho de Narukawa, Leiva e Fujita (2009) traz parte dos resultados da dissertação de Narukawa (2011), que estende a aplicação da indexação automática para artigos da área agrícola da base BINAGRI, com uso do tesauro ThesAgro. Narukawa (2011) reporta as consistências médias de 19,30% e 23,25% encontradas respectivamente na aplicação do software SISA na indexação automática dos artigos da área agrícola da base BINAGRI com uso do ThesAgro e dos artigos da área de odontologia da base BIREME com o uso do DeCS.

Outro trabalho que fez uso do SISA como instrumento de indexação foi o de Lima e Boccato (2009). As autoras avaliaram o desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP nos processos de indexação manual, automática e semiautomática de um corpus composto por resumos de dissertações e teses. Os descritores foram assim determinados: Descritores atribuídos por indexadores às teses e dissertações produzidas no Programa de Pós-Graduação em Ciência da Informação (PPGCI) no período de 2002 a 2007 e armazenadas no Banco de Dados Bibliográficos da USP – (DEDALUS); Descritores atribuídos automaticamente pelo SISA aos resumos das dissertações e teses; Descritores atribuídos de forma semiautomática com uso do SISA, isto é: descritores selecionados pelo indexador a partir dos descritores atribuídos e os termos candidatos selecionados pelo SISA como possíveis descritores.

Segundo as autoras, no contexto em que há necessidade de agilizar a indexação de fontes de informação e que os usuários estão cada vez mais exigentes na satisfação das suas necessidades de buscas por informações, são essenciais as linguagens documentárias pois propiciam a compatibilidade entre a linguagem do usuário e a utilizada pelo sistema de recuperação da informação.

Lima e Boccato (2009) apresentam o Vocabulário Controlado do SIBi/USP como uma linguagem documentária construída pelos bibliotecários, tendo por base normas e procedimentos tanto terminológicos como documentários, e que contou com a participação de especialistas de todas as áreas do conhecimento abrangidas pelos seus descritores.

Quanto à metodologia aplicada do trabalho, foram realizadas as indexações dos documentos de um corpus composto por 70 resumos de teses e dissertações do PPGCI cadastradas no Banco de Dados Bibliográficos da USP (DEDALUS), abrangendo o período de janeiro de 2002 a dezembro de 2007.

As principais conclusões apontadas pelas autoras após análise das indexações realizadas foram:

a) Em alguns casos os resumos não são elaborados adequadamente, não permitindo uma indexação automática ou semiautomática de qualidade por não incluir a terminologia do domínio no qual estão inseridos;
b) Durante a indexação semiautomática, percebeu-se que em alguns casos o indexador identificava o assunto tratado na tese ou dissertação, e o assunto fazia

parte do Vocabulário Controlado, mas não foi indexado pelo SISA, pois o mesmo não fazia parte do resumo escrito pelo autor da tese ou dissertação;

c) Os descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP precisam ser ampliados e contextualizados através de definições terminológicas, visando representarem adequadamente o conteúdo do corpus objeto dos processos de indexação. Isto pode ser observado, segundo as autoras, pela inexistência no Vocabulário Controlado SIBi/USP de novos termos, os quais aparecem nos resumos como novos assuntos do domínio, mas que não foram ainda incorporados ao Vocabulário como descritores na área da Ciência da Informação;

Através da análise do desempenho terminológico dos descritores atribuídos pelos três processos de indexação, via percentual de coincidência de termos ou termos comuns, as autoras revelam que ao comparar a indexação automática do SISA com a indexação manual dos autores: para 67% dos documentos não houve termos comuns; para 27% dos documentos houve um termo comum atribuído; e para 6% dos documentos houve dois termos comuns atribuídos para cada documento. Embora não tenha sido reportado o índice de consistência, os valores reportados indicam um baixo desempenho na indexação automática por atribuição.

Em (SOUZA; LEIVA, 2016) e (LEIVA, 2017) são reportados experimentos com aplicação do software SISA e vocabulário controlado Thesagro na indexação automática de artigos em português da área de Fruticultura. São reportadas métricas de revocação e precisão na recuperação da informação para um conjunto de expressões de busca a fim de mensurar a qualidade da indexação automática. Tais trabalhos diferem do presente artigo por utilizarem artigos e vocabulário controlado da área de fruticultura. São reportadas métricas de qualidade na recuperação da informação, mas não é reportada a consistência como métrica de qualidade da indexação automática.

Finalizando a discussão dos trabalhos relacionados, vale discutir o trabalho de Souza (2005) que propôs um método de indexação automática por sintagmas nominais, nos moldes de uma indexação automática por extração.

Em sua pesquisa de doutorado, Souza (2005) aborda a questão do uso de sintagmas nominais na representação do conteúdo informacional dos documentos. Realizada a extração automática de sintagmas nominais, o autor propõe um método de seleção de expressões que se constituem em sintagmas nominais no contexto linguístico e tenham relevância na representação dos assuntos.

Embora não aborde um caso de indexação automática por atribuição e nem tenha reportado a consistência na indexação obtida através do método proposto, utiliza-se do mesmo corpus de artigos de periódicos científicos na área de Ciência da Informação que Souza (2005), entre outros motivos por ter sido utilizado em outros trabalhos como (SOUZA, 2006), (SOUZA; RAGHAVAN, 2006) e (SOUZA; RAGHAVAN, 2014), possibilitando a posterior comparação dos resultados de diferentes métodos de indexação automática para o corpus.

Levando em conta os trabalhos descritos nessa seção, o presente trabalho contribui de forma inédita na análise da consistência na indexação automática por atribuição de artigos científicos escritos em português da área de Ciência da Informação.

Na próxima seção são descritos os materiais e métodos da presente pesquisa.

3 MATERIAL E MÉTODOS

Trata-se de uma pesquisa exploratória, tendo como finalidade, segundo Gil (1996), proporcionar uma familiarização com o problema visando construir novas hipóteses, tornando-o mais claro ou explícito.

Adotou-se procedimentos técnicos a pesquisa bibliográfica – via revisão da literatura acerca da indexação automática por atribuição de artigos científicos escritos em português – e experimento computacional – envolvendo a aplicação e avaliação da indexação automática por atribuição em corpus de artigos científicos da área de Ciência da Informação.

No experimento computacional, utilizou-se do Sistema de Indización Semi-Automático (SISA) (LEIVA, 2009) como sistema de indexação automática por atribuição. A utilização deste sistema é justificada pela sua aplicação na grande maioria dos trabalhos relacionados e por ter sido gentilmente disponibilizado pelo seu desenvolvedor para fins de pesquisa.

Utilizou-se como linguagem de indexação o Tesauro Brasileiro de Ciência da Informação (TBCI) (PINHEIRO; FERREZ, 2014), por se tratar de um tesauro atualizado, publicado recentemente, e que contempla os termos mais relevantes para representação dos assuntos das publicações científicas da área de Ciência da Informação no Brasil.

O corpus para este estudo consiste de uma amostra de 60 artigos científicos selecionados por Souza (2005). Trata-se de todos os artigos científicos escritos em língua portuguesa, publicados durante os anos de 2002 e 2003 em duas revistas científicas eletrônicas da área da Ciência da Informação. As revistas foram DataGramaZero e Ciência da Informação do IBICT, periódicos científicos reconhecidos pelo programa QUALIS da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) naqueles anos.

Utiliza-se o corpus de Souza (2005) na presente pesquisa pelos seguintes motivos: apresenta um número estatisticamente significante de artigos científicos na área de Ciência da Informação; foi utilizado em outros trabalhos como (SOUZA, 2006) (SOUZA; RAGHAVAN, 2006), (SOUZA; RAGHAVAN, 2014); possibilita a posterior comparação dos resultados de diferentes métodos de indexação automática para o corpus; como a data de publicação dos artigos antecede ao da publicação do tesauro utilizado nesta pesquisa, presume-se que os termos de indexação mais relevantes dos artigos estão contemplados no tesauro (garantia literária do TBCI).

Os 60 artigos escolhidos foram convertidos para o formato texto, uma vez que o SISA exige este formato para poder efetuar a indexação. A formatação dos artigos assemelha-se à realizada por Souza (2005), porém acrescidos de formatação específica para o SISA.

Os passos para formatação dos textos completos para entrada no software SISA foram:

1) Os arquivos originais em formato HTML foram abertos no Microsoft Word, e os arquivos em formato PDF foram abertos no Acrobat PDF Reader, sendo posteriormente salvos no formato TXT;
2) Os arquivos no formato TXT foram editados no Word de forma a manter somente o título, resumo em português e texto (da introdução às considerações finais ou conclusão), e os caracteres dos arquivos foram convertidos para maiúsculas ou caixa alta;
3) Foram marcados o início e fim do título, início e fim do resumo e início e fim do texto com os marcadores #CTI# e #FTI#, #CR# e #FR#, e #CT# e #FT# respectivamente;
4) Os arquivos foram salvos escolhendo a opção “texto sem formatação”.

Foram utilizados o Adobe Acrobat Reader V. 11.0 e o Microsoft Office Word 2013 para formatação dos arquivos.

Visando a indexação dos arquivos texto, o SISA, foi configurado com os seguintes arquivos no formato texto, sendo todos os caracteres convertidos para caixa alta:

1) Linguagem de indexação – uma lista de termos organizados em ordem alfabética e numerados sequencialmente, os quais, podem ter uma relação de equivalência e termo autorizado com outro termo por meio do termo USE. Foi criado um arquivo, extraindo conteúdo do TBCI convertido para o formato TXT;
2) Lista de palavras vazias ou stopwords – a lista de palavras vazias em língua portuguesa foi cedida pelo coordenador do projeto de pesquisa denominado Mapeador Temático de Teses e Dissertações (MTTD), e adaptada removendo as que faziam parte de algum termo da lista de descritores do TBCI, uma vez que, a sua presença, interfere na atribuição de termos de indexação. Este arquivo também foi salvo no formato TXT;
3) Artigos Científicos – cada um dos 60 arquivos de texto contendo o texto completo dos respectivos artigos científicos. O título, resumo e texto foram marcados (delimitados) por #CTI# (começo do título), #FTI# (fim do título), #CR# (começo do resumo), #FR# (fim do resumo), #CTE# (começo do texto) e #FTE# (fim do texto).

A conversão de todos os caracteres dos arquivos de entrada do SISA para caixa alta foi realizada para garantir que o software execute corretamente diante dos caracteres especiais da língua portuguesa, como os caracteres acentuados e cê-cedilha.

Posteriormente à entrada dos arquivos de configuração no SISA, foi solicitado a indexação automática de cada um dos arquivos dos artigos científicos.

Após a indexação dos 60 artigos, os termos de indexação propostos pelo SISA foram dispostos num quadro onde constam também as palavras-chave dos respectivos artigos. Assim, tornou-se possível comparar os dois conjuntos de termos de indexação, marcar os termos comuns que casam totalmente (negrito) e parcialmente (negrito e itálico), e posteriormente avaliar a consistência na indexação. O Quadro 1 ilustra uma amostra da planilha eletrônica criada para comparação dos termos de indexação.

Quadro 1: Amostra da planilha para comparação dos termos de indexação.

Artigo Científico	**Termos das Palavras-chave**	**Termos de Indexação do SISA**
artigo 1	**1-transferência de informação** **2-gestão do conhecimento** 3-valor de unidades de conhecimento.	**1-transferencia da informação** 2-avaliação 3–acesso 4–descarte 5-estudos de caso 6–gestão **7 - gestão do conhecimento**
artigo 2	1-popularização da ***ciência*** **2 - comunicação científica.**	**1-comunicação cientifica** 2-***ciencia*** da informação 3-estudos de caso 4–educação 5 – noticias
artigo 3	***1–informação*** 2-valor informacional **3-direito à informação** 4-memória social 5-estoque informacional.	1–avaliação 2–direito **3-direito a informação** 4-recuperação da ***informação***

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Posteriormente, neste quadro foram incluídas as seguintes colunas para cada artigo: (A) número de termos das palavras chaves definidas pelos autores, (B) número de termos atribuídos pelo SISA, e (Tco) correspondendo ao número de termos comuns ao se comparar os termos das duas indexações.

Para cálculo dos índices de consistência, foi aplicada a fórmula do cálculo de consistência (Equação 1). Esta fórmula foi utilizada para se obter o índice de consistência na indexação automática de cada artigo.

$$Ci=Tco/((A+B)-Tco) \qquad \textbf{(Equação 1)}$$

Onde:
Ci = Índice de consistência
Tco = Número de termos comuns nas duas indexações
A = Número de termos usados na indexação A
B = Número de termos usados na indexação B

No presente trabalho, a indexação A corresponde à realizada pelos autores dos artigos e a indexação B corresponde à indexação automática por atribuição realizada pelo software SISA utilizando o TBCI como linguagem de indexação.

Essa fórmula foi utilizada com algumas variações por diversos trabalhos para verificar a consistência entre a indexação manual e a automática (NARUKAWA, 2011).

Ao se analisar a coincidência entre os termos de indexação e aplicar os valores na fórmula de consistência, considerou-se o critério de comparação relaxada ou critério de consistência relaxada (LEIVA, 2008). Assim, quando há coincidência total entre os termos que estão sendo comparados, tem-se o valor 1 (100%), quando há coincidência parcial tem-se valor 0,5 (50%) e, quando não há coincidência, tem-se valor 0 (0%).

No presente trabalho, para se categorizar se os valores alcançados para o índice de consistência são satisfatórios faz-se uso de duas prerrogativas: a proximidade dos valores alcançados neste trabalho e nos trabalhos relacionados para sistemas de indexação automática por atribuição; e o nível de consistência médio entre 11 a 25%, o que corresponde à uma média de um a dois termos de indexação em comum por documento.

Trata-se de uma classificação relativa e sujeita a interferências de diversos fatores (área temática, vocabulário controlado, quantidade de documentos, quantidade de descritores, metodologia do experimento, etc.), mas aceitável e necessária no presente contexto.

Com base na fórmula do índice de consistência foram calculadas as seguintes faixas de valores para o índice de consistência médio correspondentes aos níveis de desempenho de um sistema de indexação automática: 0 a 11% - desempenho insatisfatório (corresponde a uma média de 0 a 1 termo em comum); 11 a 25% - desempenho satisfatório (corresponde a uma média de 1 a 2 termos em comum); 25 a 43% - desempenho bom (corresponde a uma média de 2 a 3 termos em comum); 43 a 67% - desempenho ótimo (corresponde a uma média de 3 a 4 termos em comum); 67 a 100% - desempenho excelente (corresponde a uma média de 4 a 5 termos em comum). Para cálculo dos limites dos intervalos foi estipulado o número médio de termos de indexação atribuídos pelas duas indexações iguais a cinco.

O critério de consistência relaxada foi aplicado em dois cenários: I) Cenário I - que consiste em habilitar a atribuição de termos gerais pelo SISA; II) Cenário II - que consiste em não habilitar a atribuição de termos gerais pelo SISA. Justifica-se a análise desses dois cenários devido ao impacto proporcionado pela atribuição de termos gerais nos índices de consistência na indexação.

4 RESULTADOS E DISCUSSÃO

Nesta seção apresenta-se a análise da consistência na indexação automática por atribuição, tomando como padrão de referência a indexação intelectual dos autores dos trabalhos.

Para os dois cenários do experimento, o Quadro 2 mostra a média e o desvio padrão dos seguintes indicadores para o corpus: o número de palavras-chave, número de termos propostos pelo SISA, número de termos comuns entre as palavras-chave e os atribuídos pelo SISA, e o percentual do índice de consistência. Os resultados foram obtidos com o arquivo de termos gerais habilitado (Cenário I) e com o arquivo desabilitado (Cenário II).

Comparando os dois cenários, a média do número de termos do SISA atribuídos a cada artigo no Cenário I é estatisticamente superior à média do Cenário II, enquanto que a diferença nas médias do número de palavras-chave comuns não é estatisticamente significante.

Quadro 2: Médias e desvios padrões dos indicadores nos dois cenários.

Cenário	No. de Palavras-chave	No. de termos do SISA	No. de Termos comuns	Consistência Relaxada
Cenário I	4,5 ± 1,6	9,9 ± 4,9	1,7 ± 1,0	14% ± 8%
Cenário II	4,5 ± 1,6	5,1 ± 2,1	1,5 ± 1,0	19% ± 13%

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Consequentemente, a média do índice de consistência no Cenário II é estaticamente superior à do Cenário I. Tal média superior pode ser explicada pela redução aproximadamente pela metade da média de termos propostos pelo SISA (ao não propor termos gerais), e a quase manutenção da média do número de termos comuns com as palavras-chave.

É importante ressaltar que os valores médios de consistência de 14% e 19% estão dentro do patamar encontrado nas pesquisas sobre consistência na indexação, cujos estudos revelam que os mesmos variam entre 10% e 60% (NARUKAWA, 2011).

Além disso, a média encontrada para o índice de consistência de 19% é muito próxima das médias de 19,30% e 23,25%, encontradas por Narukawa (2011), na aplicação do software SISA para indexar, respectivamente, os artigos da área agrícola da base BINAGRI com uso do ThesAgro e dos artigos da área de odontologia da base BIREME com o uso do DeCS.

Embora em domínios diferentes, estas consistências médias parecem indicar um patamar dos valores médios do índice de consistência da indexação automática por atribuição utilizando o SISA (com o uso de um tesauro de especialidade), tomando as palavras-chave dos autores como padrão de referência na indexação de artigos científicos.

Levando em conta que o índice de consistência médio de 19% se aproxima dos valores reportados na literatura e que está na faixa de valores entre 11 e 25%, pode-se categorizar o desempenho como satisfatório na indexação automática por atribuição do presente corpus.

Adicionalmente, o número médio de palavras-chave comuns de uma ou duas palavras-chave por artigo, encontradas em ambos os cenários, é semelhante ao resultado obtido por Miguéis et tal (2013) como a média de palavras-chave equivalentes por artigo encontradas na comparação das palavras-chave dos autores com as atribuídas pelos indexadores na base de dados MEDLINE. Embora os domínios de especialidade sejam diferentes, este fato comum pode ser um indício da proximidade da indexação automática por atribuição e a indexação intelectual, tendo as palavras-chave dos autores como padrão de referência.

O número médio de palavras-chave comuns por documento obtido de 1,5 e 1,7 termos comuns supera o 0,4 obtido em (LIMA; BOCCATO, 2009) para a indexação automática de teses e dissertações em Ciência da Informação utilizando o SISA com o vocabulário controlado do SIBi/USP. Atribui-se como possíveis causas deste melhor desempenho, a maior abrangência do TBCI na representação da terminologia da área, o uso do critério de consistência relaxada no presente trabalho, e pelo fato do SISA ter sido desenvolvido para indexar textos completos de artigos científicos.

Apresenta-se a seguir, para os dois cenários do experimento, uma análise dos casos mais consistentes e casos menos consistentes com base nos valores mais altos e mais baixos do índice de consistência alcançados pelo SISA.

Os seis casos mais consistentes e os seis casos menos consistentes no Cenário I são descritos na Tabela 1 e Tabela 2 respectivamente.

Para os casos mais consistentes no Cenário I, percebe-se que o índice de consistência variou de 27% a 42%. Esses melhores valores de consistência foram determinados por um bom número de termos comuns, que variou de 1,5 a 5 descritores comuns, e um número de termos propostos pelo SISA em geral próximos do número de palavras-chave.

Tabela 1: Os seis casos mais consistentes da indexação feita pelo SISA com a atribuição de termos gerais habilitada (Cenário I).

Artigo Científico	No. de Palavras-chave	No. de termos do SISA	No. de Termos comuns	Consistência Relaxada
Artigo 9	5	4	2	29%
Artigo 15	2	5	1,5	27%
Artigo 16	7	7	3,5	33%
Artigo 29	4	9	3	30%
Artigo 44	3	6	2	29%
Artigo 58	7	10	5	42%

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Tabela 2: Os seis casos menos consistentes da indexação feita pelo SISA com a atribuição de termos gerais habilitada. (Cenário I).

Artigo Científico	No. de Palavras-chave	No. de termos do SISA	No. de Termos comuns	Consistência Relaxada
Artigo 5	5	3	0	0%
Artigo 10	5	2	0	0%
Artigo 19	5	5	0	0%
Artigo 23	3	1	0	0%
Artigo 31	6	3	0	0%

Artigo 34	3	3	0	0%

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Para os casos menos consistentes apresentados na Tabela 2, o índice de consistência variou de 0% a 6%, sendo determinado praticamente pelo número baixo de termos comuns, que variou de 0 a 0,5 termos, enquanto o número de termos propostos pelo SISA se manteve em geral próximo ao número de palavras-chave na maioria dos casos.

Para a indexação automática por atribuição pelo SISA com a atribuição de termos gerais desabilitada (Cenário II), a Tabela 3 exibe os sete casos mais consistentes e a Tabela 4 exibe os seis casos menos consistentes.

Para os casos mais consistentes no Cenário II, percebe-se que o índice de consistência variou de 30% a 71%. Esses melhores valores de consistência foram determinados por um bom número de termos comuns, que variou de 1,5 a 5 descritores comuns, e um número de termos propostos pelo SISA mais próximos do número de palavras-chave.

Tabela 3: Os sete casos mais consistentes da indexação feita pelo SISA com a atribuição de termos gerais desabilitada (Cenário II).

Artigo Científico	**No. de Palavras-chave**	**No. de termos do SISA**	**No. de Termos comuns**	**Consistência Relaxada**
Artigo 15	2	3	1,5	43%
Artigo 16	7	4	3,5	47%
Artigo 29	4	5	3	50%
Artigo 35	5	8	3	30%
Artigo 39	5	8	3	30%
Artigo 58	7	5	5	71%

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Para os casos menos consistentes, o índice de consistência foi 0% para todos os casos, sendo praticamente determinado pelo número de termos comuns igual a zero, enquanto o número de termos propostos pelo SISA se manteve menor ou igual ao número de palavras-chave.

Comparando os percentuais obtidos nos dois cenários para os casos mais consistentes, através das Tabelas 1 e 3, verifica-se que os índices de consistência foram alterados. Mais uma vez verifica-se a diminuição dos valores dos índices de consistência quando a indexação é feita habilitando-se a atribuição de termos gerais.

Tabela 4: Os seis casos menos consistentes da indexação feita pelo SISA com a atribuição de termos gerais desabilitada (Cenário II).

Artigo Científico	**No. de Palavras-chave**	**No. de termos do SISA**	**No. de Termos comuns**	**Consistência Relaxada**
Artigo 5	5	3	0	0%
Artigo 10	5	2	0	0%
Artigo 19	5	5	0	0%
Artigo 23	3	1	0	0%
Artigo 31	6	3	0	0%
Artigo 34	3	3	0	0%

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Analisando-se os casos mais consistentes comuns aos dois cenários, vide Tabela 5, observa-se que a inclusão dos termos gerais diminui os valores percentuais dos índices de consistência, isto porque o número de termos comuns permaneceu o mesmo e o número de termos atribuídos pelo SISA foi maior. Analisando-se o artigo 15, que aparece entre os casos mais consistentes nos dois cenários, percebe-se que o índice de consistência passou de 27% no Cenário I para 43% no Cenário II. Isto porque o número de termos atribuídos pelo SISA foi

menor no Cenário II (três termos) do que no Cenário I (cinco termos), sendo esse padrão válido também para os demais artigos (16, 29 e 58).

Tabela 5: Análise do índice de consistência dos casos mais consistentes comuns aos Cenários I e II

Artigos	Índices de Consistência Cenário I	Índices de Consistência Cenário II	No. De Termos comuns Cenários I e II	No. De Termos SISA Cenário I	No. de Termos SISA Cenário II
15	27 %	43 %	1,5	5	3
16	33 %	47 %	3,5	7	4
29	30 %	50 %	3	9	5
58	42%	71%	5	10	5

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Analisando-se os casos menos consistentes comuns aos dois cenários a partir da Tabela 6, verifica-se a diminuição dos valores percentuais dos índices de consistência.

Por exemplo para o artigo 10, o índice de consistência obtido no Cenário I foi de 5% e no Cenário II foi de 0%.

Tabela 6: Análise do índice de consistência dos casos menos consistentes comuns aos Cenários I e II

Artigos	Índices de Consistência Cenário I	Índices de Consistência Cenário II	No. De Termos comuns Cenário I	No. De Termos comuns Cenário II	No. De Termos SISA Cenário I	No. De Termos SISA Cenário II
5	6%	0%	0,5	0	4	3
10	5%	0%	0,5	0	5	2
23	0%	0%	0	0	3	1
31	4%	0%	0,5	0	7	3
34	0%	0%	0	0	4	3

Fonte: dados do autor a partir da análise do corpus e da saída do software SISA

Os valores da Tabela 6 indicam que a habilitação dos termos gerais (Cenário I), produziu uma pequena melhora no índice de consistência para a maioria dos casos menos consistentes, isto porque, a atribuição dos termos gerais possibilitou um número de termos comuns com as palavras-chave igual a 0,5 usando o critério de consistência relaxada. Este 0,5 termo comum corresponde, na maioria absoluta dos casos, a uma palavra do termo geral atribuído equivalente a uma palavra de uma palavra-chave, por instância: informação; e científica.

Entretanto, para que este 0,5 termo comum fosse encontrado, foram atribuídos pelo SISA muitos termos gerais aos documentos que não eram equivalentes às palavras-chave, levando a um índice de consistência, em geral mais baixo para os demais documentos. De qualquer forma, em apenas 10% dos casos (6 documentos) foram obtidos valores de consistência iguais a 0% para o Cenário II, e esses valores se encontram muito abaixo do percentual médio do índice de consistência obtido na indexação dos 60 artigos que foi de 19%.

Portanto, levando-se em conta a indexação de todo o corpus, a habilitação da atribuição de termos gerais no SISA gerou piores resultados em termos de consistência na indexação automática por atribuição, medida pelo índice de consistência na comparação com a indexação intelectual dos autores dos artigos.

5 CONSIDERAÇÕES FINAIS

A utilização da indexação automática por atribuição com base no Tesauro Brasileiro de Ciência da Informação (TBCI) e o uso do software SISA, proporciona uma consistência média satisfatória na indexação de artigos científicos em português da área de Ciência da Informação. Tal conclusão se baseia no fato que o percentual médio obtido para o índice de consistência de 19% se encontra próximo dos valores reportados na literatura científica para indexação automática por atribuição de artigos de outras áreas do conhecimento, bem como

se encontra na faixa de valores do índice de consistência médio correspondente ao desempenho satisfatório.

Este valor médio de consistência pode ser interpretado como o comportamento médio de 19% de termos comuns do total de termos do conjunto união dos termos da indexação intelectual dos autores e da indexação automática por atribuição. Ou seja, para um caso médio típico tem-se uma a duas palavras-chave comuns por documento, dado cinco palavras-chave dos autores e cinco termos atribuídos pelo software. Trata-se de um desempenho satisfatório, com margem para melhorias até os 25% neste mesmo patamar, a partir do qual entende-se que consistiria de um desempenho bom (duas a três palavras-chave comuns por documento).

Entretanto, para alcance do patamar médio de 19% de consistência na indexação automática por atribuição utilizando o software SISA foi desabilitada a atribuição de termos gerais. Isto porque a atribuição dos termos gerais pelo SISA resulta em um número elevado de termos propostos para indexação automática, que quando comparados com os termos das palavras-chave em sua maioria não coincidem, resultando em baixos índices de consistência.

Através da análise dos resultados reportados neste trabalho, pode-se concluir também que a obtenção de consistência mais alta na indexação automática por atribuição depende primeiramente da atribuição de termos presentes no vocabulário controlado como palavras-chave dos artigos, bem como da presença destes termos também no conteúdo textual dos documentos. Assim, as palavras-chave são determinantes para obtenção de valores mais altos dos índices na recuperação da informação e na avaliação da consistência na indexação automática.

Como limitação do corpus utilizado neste trabalho, verificou-se que entre os 60 artigos analisados existem alguns com apenas duas palavras-chave e outros com mais de oito palavras-chave. Outra limitação é que apenas cerca de 33% das palavras-chave se constituem termos do TBCI. Tais limitações podem ser minimizadas reindexando intelectualmente os trabalhos com o uso do TBCI, atribuindo um número fixo de palavras-chave para cada artigo.

Objetivando avanços nas pesquisas sobre indexação automática por atribuição, alguns trabalhos futuros podem ser elencados, como:

a) Avaliar a qualidade da indexação automática por atribuição no presente corpus através da métricas de revocação e precisão;
b) Avaliar a consistência na indexação automática por extração realizada por Souza (2005) nos artigos do corpus, e comparar com a obtida neste trabalho;
c) Reindexar os documentos do corpus utilizando o TBCI como linguagem de indexação e, reavaliar a consistência na indexação automática por atribuição, levando em conta os termos dos indexadores como padrão de referência ao invés das palavras-chave dos autores;
d) Pesquisar por mudanças no algoritmo do SISA, para que o mesmo não atribua uma grande quantidade de termos gerais sem potencial aumento da consistência na indexação automática por atribuição;
e) Desenvolver instrumentos que deem suporte aos autores de artigos científicos na área de Ciência da Informação, na definição das palavras-chave de seus trabalhos utilizando termos do TBCI.

REFERÊNCIAS

CORRÊA, R. F.; LAPA, R. C. Panorama de Estudos sobre indexação automática no âmbito da Ciência da Informação no Brasil (1973-2012). **Ciência da Informação**. Brasília, DF, v. 42, n. 2, p. 255-273, 2013. Disponível em: <http://revista.ibict.br/ciinf/article/view/1385>

FUJITA, M. S. L. A representação documentária de artigos científicos em educação especial: orientação aos autores para determinação de palavras chaves. **Revista Brasileira de Educação Especial**, Marília, v. 10, n. 3, p.257-272, 2004.

GIL, A. C. **Como elaborar projetos de pesquisa**. 3. ed. São Paulo: Atlas, 1996. 159 p.

GUINCHAT, C.; MENOU, M. **Introdução geral às Ciências e técnicas da informação e documentação**. 2. ed. corr. aum. Brasília: IBICT, 1994.

KOBASHI, N. Y. **A elaboração de informações documentárias: em busca de uma metodologia**.1994. 195 f. Tese (Doutorado em Ciências da Comunicação), Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo.

LAPA, R.; CORRÊA, R. F. Indexação Automática no Âmbito da Ciência da Informação no Brasil. **Informação & Tecnologia (ITEC)**. Marília/João Pessoa, v. 2, n. 1, p.1-18, 2014.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. 2. ed. ver. atual. Brasília: Briquet de. Lemos, 2004.

LEIVA, I. G. **La automatización de la indización de documentos**. Gijón: Trea, 1999.

LEIVA, I. G., Consistencia en la asignación de materias en bibliotecas públicas del Estado**. Boletín de la Asociación Andaluza de Bibliotecarios,** n. 63, p. 69-96, 2001.

LEIVA, I. G., **Manual de Indizacíon: Teoría y Prática**. Gijón: Trea, 2008. 429p.

LEIVA, I. G. SISA – Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules". **Knowledge Organization**, v.44, n. 3, p. 139-162, 2017.

LIMA, V. N. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em ciência da informação do vocabulário controlado do SIBi/usp nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362009000100010&lng=en&nrm=iso&tlng=pt>

MIGUÉIS, A. et al. A importância das palavras-chave dos artigos científicos da área das Ciências Farmacêuticas, depositados no Estudo Geral: estudo comparativo com os termos atribuídos na MEDLINE. InCID: **Revista de Ciência, Informação e Documentação**, São Paulo, v. 4, n. 2, p.112-125, 2013. Disponível em: <http://www.revistas.usp.br/incid/article/view/69284>

NARUKAWA, C. M.; LEIVA, I. G.; 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**, João Pessoa, v. 19, n. 2, 2009. Disponível em: < http://www.periodicos.ufpb.br/ojs/index.php/ies/article/view/2925>

NARUKAWA, C. M. **Estudo de Vocabulário Controlado na Indexação Automática: Aplicação no Processo de Indexação do Sistema de Indización Semiautomatica (SISA)**. 222 f. Dissertação (Mestrado em Ciência da Informação), Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro, Brasília: IBICT, 2014. 384 p.

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 DF: Edição de autor, 2005. 410 p.

SOUZA, R. R**. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.** 215 f. Tese (Doutorado em Ciência da Informação), Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 11, n. esp., p. 42-59, 2006. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2006v11nesp1p42>.

SOUZA, R. R.; LEIVA, I. G. Automatic Indexing of Scientific Texts: A Methodological Comparison. In: INTERNATIONAL SOCIETY OF KNOWLEDGE ORGANIZATION COONFERENCE, 14, Rio de Janeiro. **Anais...** Würzburg: Ergon Verlag, p. 243-250, 2016.

SOUZA, R. R.; RAGHAVAN, K. S. A methodology for noun phrase-based automatic indexing. **Knowledge Organization**, v. 33, n. 1, p. 45-56, 2006.

SOUZA, R. R.; RAGHAVAN, K. S. Extraction of keywords from texts: an exploratory study using Noun Phrases. **Informação & Tecnologia (ITEC)**. Marília/ João Pessoa. v. 1, n. 1. p. 5-16, 2014.

Editores do artigo: Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.