XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT 11 – Informação e Saúde

USO DA INDEXAÇÃO AUTOMÁTICA NA REPRESENTAÇÃO DE ARTIGOS EM PORTUGUÊS DA ÁREA DE SAÚDE PÚBLICA

USE OF AUTOMATIC INDEXING IN THE REPRESENTATION OF ARTICLES IN PORTUGUESE IN THE PUBLIC HEALTH AREA

Fatima Cristina Lopes dos Santos. FIOCRUZ.

Cícera Henrique da Silva. FIOCRUZ.

Rosane Abdala Lins. FIOCRUZ.

Modalidade: Trabalho Completo

Resumo: Este estudo verifica o grau de coerência semântica entre a representação temática via linguagem controlada e a indexação automática dos artigos publicados em português na área de saúde pública, com o aporte teórico da mineração de textos. Propõe-se a análise da coerência semântica entre o conteúdo dos trabalhos e sua representatividade (palavras-chave e descritores), observando o uso da indexação automática e da linguagem controlada na representação temática dessa informação em saúde. Utilizando artigos publicados em português no periódico Cadernos de Saúde Pública, integrante da Coleção Saúde Pública na SciELO Brasil, viu-se que a mineração de texto e a indexação automática pode colaborar com a representação do conteúdo dos artigos estudados. Compreendendo a indexação como representação de conteúdo e seu importante papel no sistema de comunicação na ciência, procura-se refletir sobre a informação em saúde representada nesses textos, entendendo a relevância da visibilidade e recuperação dessa produção. Vale ressaltar que o periódico científico - e por conseguinte seus artigos - além de outros papéis estratégicos na comunicação da ciência, é registro da memória, chancela de qualidade e difusores do conhecimento. Espera-se que este estudo contribua com as pesquisas sobre a indexação de artigos científicos na área de saúde pública, particularmente na coerência semântica da representação de artigos publicados em português e o uso da indexação automática.

Palavras-Chave: Indexação Automática. Mineração de Textos. Coerência Semântica. Artigos Científicos. Saúde Pública.

Abstract: This study verifies the degree of semantic coherence between thematic representation via controlled language and the automatic indexing of articles published in Portuguese in the field of public health, with the theoretical support of text mining. The analysis of the semantic coherence between the content of the works and their representativeness (keywords and descriptors) is proposed, observing the use of automatic indexing and controlled language in the thematic representation of this health information. Using articles published in Portuguese in the journal Reports in Public Health, part of the Public Health Collection at SciELO Brasil, it was seen that text mining and automatic indexing could collaborate with the representation of the content of the studied articles. Understanding

indexing as a representation of content and its important role in the communication system in science, we seek to reflect on the health information represented in these texts, understanding the relevance of the visibility and recovery of this production. It is noteworthy that the scientific journal - and therefore its articles -, in addition to other strategic roles in the communication of science, are memory records, quality seals and knowledge diffusers. It is expected that this study will contribute to research on the indexing of scientific articles in the field of public health, particularly in the semantic coherence of the representation of articles published in Portuguese and the use of automatic indexing.

Keywords: Automatic Indexing. Text Mining. Semantic Coherence. Scientific Articles. Public Health

1 INTRODUÇÃO

Para obtenção de alto grau de precisão na recuperação da informação na web - além da especificidade da linguagem de indexação - faz-se necessária a coerência na representação da informação, visando a busca orientada em ambientes virtuais organizados para tal. Assim, pode-se associar o grau de coerência da indexação - extensão com que há concordância quanto aos termos a serem usados para indexar um documento (LANCASTER, 2004) - com o grau de precisão de um sistema de recuperação da informação, isto é, o desempenho desse sistema quando a maioria dos itens recuperados é considerada útil ou inútil por quem realiza ou solicita a busca nesse ambiente (ARAÚJO JÚNIOR, 2007). Quando esses meios virtuais tratam de informação em saúde e do processo de comunicação científica, a representação da informação envolve o complexo modelo de negócio de periódicos científicos - ainda o principal canal de comunicação e divulgação entre pares. Sobre o papel do artigo de periódico na comunicação científica, Curty e Boccato consideram que:

> A comunicação científica é um processo complexo de produção, disseminação e uso adequado da informação. Para que essa comunicação desempenhe seu papel de promotora da informação científica, o artigo científico, como um dos meios dessa comunicação, deve mostrar-se de forma objetiva, coerente, concisa, com propriedade sintática, clareza semântica e padronizada. Essa padronização seguirá normas estabelecidas de organização e normalização da informação quanto a sua forma, estrutura, linguagem e conteúdo (CURTY; BOCCATO, 2005, p.106).

Dado que uma das funções do periódico científico é ser um canal formal de comunicação na ciência, a coerência semântica na representação de artigos científicos contribui para a organização da informação, unificando padrões na representação temática. A criação e circulação de terminologias em distintos cenários comunicativos são testemunhos de que essas cumprem, prioritariamente, a dupla função de fixar o conhecimento técnico-científico e de promover sua transferência de modo pontual (KRIEGER; FINATTO, 2004).

Sob esse ponto de vista, independentemente do processo de indexação realizado - em linguagem natural ou controlada - o autor ao construir seu texto, seleciona da terminologia de seu campo de conhecimento, termos/conceitos que ele julga adequados para explicitação e comunicação de suas ideias e reflexões. Ele molda seu discurso de acordo com seus propósitos comunicativos, e o representa tematicamente por palavras-chave, com o objetivo de chamar a atenção do leitor para os conceitos que ele considera mais importantes em sua produção (FÓRIS, 2013).

Cabe ressaltar que a indexação se apresenta de várias formas no contexto da web: enquanto ação de constituição de índices, como processo de atribuição de termos a um documento, e no universo das publicações científicas como os artigos são incluídos em bases de dados (OLIVEIRA et al, 2020). Assim, este estudo buscará evidenciar a importância da coerência semântica na indexação da informação em saúde, especificamente em artigos na área de saúde pública disponíveis em português. Nesta pesquisa o foco será a indexação como processo de atribuição de termos para representação de artigos da saúde pública, a fim de contribuir com os estudos sobre a indexação automática - seleção/extração de termos como elementos descritivos de um documento pelo processamento de seu conteúdo com o uso de programas de computador (LANCASTER, 2004; LAPA; CORREA, 2014; GIL-LEIVA, 2017) - e particularmente no que se refere à representação temática de artigos publicados no periódico Cadernos de Saúde Pública, integrante da coleção SciELO Saúde Pública.

Propõe-se a investigação da representação temática de artigos publicados em português na área de saúde pública, por meio da análise da coerência semântica entre o conteúdo dos trabalhos e sua representação, isto é, de que forma o conteúdo desses textos se reflete em suas palavras-chave/descritores e nas palavras/termos identificados via mineração de textos.

2 FUNDAMENTAÇÃO TEÓRICA

O autor de um texto científico visa que sua produção seja identificada em sua comunidade como produto da ciência certificada, que ao ser avaliado e aceito por seus pares passe a incorporar o conhecimento produzido naquela área do conhecimento. A divulgação dos resultados de pesquisa à comunidade científica e sua posterior aprovação por essa comunidade é o que assegura visibilidade e credibilidade aos resultados encontrados (LATOUR; WOOLGAR, 1997; MEADOWS, 1999). Ao publicar seu trabalho em um periódico científico, a possibilidade

de recuperação desse texto aumenta quanto melhor seu conteúdo estiver representado. E a representação do conteúdo de um documento está na base do conceito de indexação, bem como seus conceitos relacionados: descritor, linguagem de indexação e termo de indexação. A indexação busca identificar os assuntos contidos no texto de um documento, traduzindo-os para uma linguagem de indexação que pode ser natural (palavras que ocorrem no texto) ou controlada (termos/descritores autorizados em vocabulário controlado). Quanto mais específica a representação das informações identificadas no texto em análise, maior a probabilidade de recuperação de informações relevantes pelo sistema (LANCASTER, 2004; ARAÚJO JÚNIOR, 2007).

Na indexação, termos e descritores podem ser vistos como elementos de representação e disseminação do conhecimento, mas distinguem-se pela sua natureza e características básicas. O termo é uma unidade léxica que assume valor semântico próprio em uma área de conhecimento, pela razão de integrar uma comunicação especializada. Já o descritor é um componente de uma linguagem constituída por gestores de informação, sendo, portanto, artificial (KRIEGER; FINATTO, 2004). Essa etapa de tradução de um documento, que envolve a conversão da análise conceitual num conjunto de termos de indexação, é de suma importância para a representação e está relacionada com a escolha de palavras-chave e de termos autorizados em um vocabulário controlado (LANCASTER, 2004).

As linguagens artificiais auxiliam as atividades de representação de conteúdos informacionais, viabilizando um armazenamento de informação que poderá ser facilmente recuperável. O uso de uma linguagem de indexação controlada permite a representação na indexação e a pesquisa por assunto com maior exatidão e equidade com as necessidades informacionais dos usuários (SALES, 2007; BOCATTO; TORQUETTI, 2012).

Lancaster (2004) acrescenta que a coerência na indexação se refere à extensão com que há concordância quanto aos termos a serem usados para indexar o documento. Esse autor assegura que quanto mais termos retirados do próprio conteúdo do texto, para indexação do documento, maior será a probabilidade de uma representação de qualidade. Essa qualidade influi diretamente na tarefa de recuperação da informação, se os descritores selecionados para representar o conteúdo de um documento não forem coerentes, certamente não será recuperado com facilidade, comprometendo o processo como um todo (ARAÚJO JÚNIOR, 2007; DIAS; NAVES, 2013). Assim, a coerência semântica é de suma importância na representação de

um resultante de pesquisa cientifica, que necessita de comunicação para reivindicar sua importância em seu campo de estudo. Dentre as ações que influenciam no acesso aos artigos de periódicos, a indexação é fundamental, pois identifica e atribui conceitos pertinentes ao conteúdo do documento, facilitando a precisão na sua recuperação.

Este estudo parte da necessidade de uma investigação mais aprofundada, principalmente se tratando de artigos publicados em português na área de saúde pública e sua importância na comunicação científica nacional. O Brasil passou de 17º em 2000 para 11º em 2018 entre os países que mais publicam artigos e o índice de citações do Brasil concentra quase 50% do impacto da região da América Latina e Caribe, reforçando a posição do país como líder científico dessa região (SANTIN; CAREGNATO, 2019; FAPESP, 2020). Esses dados também podem refletir que mesmo com mais de 300 anos e a introdução de novos veículos na comunicação cientifica - a relevância do periódico científico e seu formato de publicação em artigos não foram alterados. Alguns problemas são de fato comuns aos impressos e eletrônicos, como a preservação, representação e recuperação de seus conteúdos.

A partir desta problematização, a questão que se coloca é como a indexação automática - isto é, a atribuição de termos via colaboração de programas de computador - e a análise da coerência semântica podem contribuir na representação temática dessa informação em saúde. Espera-se que este estudo contribua com as pesquisas sobre a indexação de artigos científicos na área de saúde pública e o uso da indexação automática, pois a informação deve estar organizada e representada o mais fielmente possível ao seu conteúdo, sendo necessária a mediação entre o conteúdo informacional do documento e aquele que dele necessita (LANCASTER, 2004; TONELLO; LUNARDELLI; ALMEIDA JÚNIOR, 2012). Nesse contexto, pode-se tomar como exemplo a pandemia de COVID-19 e o aumento nas buscas por informações na literatura científica e o crescimento na produção de artigos sobre o assunto: a publicação de artigos científicos ao longo do curso da pandemia é essencial para apoiar novas pesquisas, além de apoiar decisões clínicas, de gestão pública e de vigilância em saúde, incluindo a vigilância sanitária (MOTA; FERREIRA; LEAL, 2020).

3 METODOLOGIA

Este estudo se baseia na Bibliometria, abordando o Processamento da Linguagem Natural (PLN) com o uso de software para mineração de textos, o que possibilitará a compreensão da forma, estrutura, volume e representação desses textos para a indexação automática dessa produção. Pelo tipo de abordagem é um estudo exploratório-descritivo composto por análise quanti-qualitativa, que reúne diferentes pontos de vista e formas de coletar e analisar os dados, permitindo uma ideia mais ampla e inteligível da complexidade da pesquisa.

O PLN é uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular textos ou falas em linguagem natural para uma variedade de tarefas (LIDDY, 2021; PUERTA-DÍAZ et al, 2021) e se divide em sete níveis:

1) Fonológico: com a interpretação de sons da fala;
2) Morfológico: busca interpretar a natureza componencial das palavras que são compostas de morfemas;
3) Léxico: busca interpretar o significado das palavras individuais;
4) Sintático: com a descoberta das estruturas gramaticais das sentenças;
5) Semântico: determina os significados das frases, concentrando-se nos significados no nível das palavras;
6) Discursivo: foca nas propriedades dos textos como um todo e faz conexões entre as frases;
7) Pragmático: busca compreender o uso intencional do idioma em situações (LIDDY, 2021).

Considera-se que esta pesquisa pode se caracterizar entre os níveis semântico e discursivo do PLN, pois utilizará a mineração de textos para extração e análise de dados a partir de textos, frases ou apenas palavras em grandes volumes de produções textuais, analisando a coerência semântica em seus conteúdos para a indexação dessas produções (LANCASTER, 2004; ARAÚJO JUNIOR, 2007).

A indexação é parte fundamental deste estudo, visto que os textos e indicadores de representatividade serão analisados levando-se em consideração as três formas de indexação de documentos citadas por Lancaster (2004): a manual, desenvolvida por humanos; a automática feita por computadores; e a híbrida, que utiliza as duas técnicas. Neste estudo, será utilizada a indexação híbrida, pois com o auxílio da mineração de textos serão comparados os termos coletados do conteúdo dos artigos com suas respectivas palavras-chave.

Assim, a partir da pergunta de pesquisa "como os artigos científicos publicados em português são representados em periódicos da área de saúde pública?", o estudo seguiu quatro etapas: 1. Seleção da fonte de informação; 2. Coleta da informação; 3. Tratamento e análise e dos dados coletados; 4. Mapeamento da atribuição de termos nos artigos científicos estudados, observando o papel da indexação automática e do grau de coerência semântica nesse processo. A seguir apresentam-se as etapas:

1. Seleção da fonte de informação: Nesta pesquisa foi selecionado o periódico Cadernos de Saúde Pública, que integra a coleção Scielo Saúde Pública. É publicado desde 1985 pela Escola Nacional de Saúde Pública Sergio Arouca, unidade da Fundação Oswaldo Cruz, e destina-se à publicação de artigos originais no campo da saúde pública, sendo reconhecido como uma das principais fontes de informação da área científica em saúde pública editada na América Latina. É uma publicação mensal de acesso aberto - com o qualis A2 no quadriênio 2013-2016 na área de saúde coletiva (PLATAFORMA SUCUPIRA, 2022), com prévia de qualis A3 no quadriênio 2017-2020 (UNB, 2022) - que informa em sua orientação aos autores que as palavras-chave dos artigos submetidos para avaliação devem constar no vocabulário controlado da área: o Descritores em Ciências da Saúde-DeCS (SciELO SAÚDE PÚBLICA, 2021; DECS, 2021).

2. Coleta da informação: O objeto de análise deste estudo é o artigo escrito em português e publicado no periódico Cadernos de Saúde Pública. Neste estudo exploratório, foram analisados os 8 artigos publicados no seu número 6 do volume 37 no ano de 2021.

3. Tratamento e análise e dos dados coletados: As palavras-chave dos artigos foram identificadas, verificando se as mesmas fazem parte, ou não, do vocabulário controlado DeCS. A mineração de textos foi realizada na versão PDF dos artigos, utilizando o programa Sobek, desenvolvido pelo Programa de Pós-Graduação em Informática na Educação/Grupo de Pesquisa GTech.Edu da Universidade Federal do Rio Grande do Sul (UFRGS, 2021). O programa foi escolhido por processar arquivos em PDF, e ser criado pensando no processamento da língua portuguesa, o que pode facilitar a análise dos textos.

4. Mapeamento da atribuição de termos nos artigos científicos estudados, observando o papel da indexação automática e do grau de coerência semântica nesse processo: A coerência semântica na representação temática dos artigos foi investigada a partir dos 3 resultados coletados nos textos: nas palavras-chave, nos descritores DeCS, e no resultado da mineração de textos.

Dessa forma buscou-se verificar se o uso da indexação automática pode colaborar com a representação temática desses textos apresentados na SciELO Saúde Pública e, principalmente, sobre o uso da indexação automática e da linguagem controlada na representação temática dessa produção em saúde pública, sendo a base para a elaboração de uma possível metodologia ou recomendação que auxilie esse processo.

Os artigos analisados foram os seguintes:

1.CARVALHO, Jamille M. Rodrigues; MONTEIRO, Simone Souza. Visões e práticas de mulheres vivendo com HIV/aids sobre reprodução, sexualidade e direitos. Cad. Saúde Pública, v.37, n.6, e00169720, 2021. Disponível em: <doi: 10.1590/0102-311X00169720>.

2.FRANÇA, Karla E. Ximenes et al. Near miss neonatal em hospitais de referência para gestação e parto de alto risco: estudo transversal. Cad. Saúde Pública, v.37, n.6, e00196220, 2021. Disponível em: <doi: 10.1590/0102-311X00196220>.

3.LIMA, Mayara Maia; COSTA, Veruska Maia; PALMEIRA, Swamy Lima; Castro, André P. Barbosa. Estratificação de territórios prioritários para vigilância da doença de Chagas crônica: análise multicritério para tomada de decisão em saúde. Cad. Saúde Pública, v.37, n.6, e00175920, 2021. Disponível em: <doi: 10.1590/0102-311X00175920>.

4.MERÊNCIO, Ivan; MONTEIRO, Gecielli Martins; VIEIRA, Carlos A. Oliveira. Aglomerados ativos de COVID-19 em Santa Catarina, Brasil, e tendência de mobilidade dos locais de trabalho. Cad. Saúde Pública, v.37, n.6, e00301620, 2021. Disponível em: <doi: 10.1590/0102-311X00301620>.

5.RODRIGUES, Marcelle N. Gomes; VIEIRA, Fernanda M.S. Barbeiro; VASCONCELLOS-SILVA, Paulo Roberto. Análise das recomendações das Comissões Regionais de Mortalidade Materna para os casos de óbitos por aborto provocado no Município do Rio de Janeiro, Brasil. Cad. Saúde Pública, v.37, n.6, e00215020, 2021. Disponível em: <doi: 10.1590/0102-311X00215020>.

6.RUIVO, Ana Carolina Oliveira et al. Disponibilidade de insumos para o planejamento reprodutivo nos três ciclos do Programa de Melhoria do Acesso e da Qualidade da Atenção Básica: 2012, 2014 e 2018. Cad. Saúde Pública, v.37, n.6, e00123220, 2021. Disponível em: <doi: 10.1590/0102-311X00123220>.

7.SANTOS, Francisca Maria Rodrigues et al. Prevalência e fatores associados a não inscrição para transplante renal. Cad. Saúde Pública, v.37, n.6, e00043620, 2021. Disponível em: <doi: 10.1590/0102-311X00043620>.

8.ZACHARIAS, Fabiana Costa Machado et al. e-SUS Atenção Primária: atributos determinantes para adoção e uso de uma inovação tecnológica. Cad. Saúde Pública, v.37, e00219520, 2021. Disponível em: <doi: 10.1590/0102-311X00219520>.

4 RESULTADOS

Verificou-se que a maioria das palavras-chave dos artigos faz parte do vocabulário controlado da área (DeCS), alcançando 100% de uso do instrumento nos artigos 2, 3, 4, 5, 7 e 8, como visto no quadro 1, o que significa que os autores desses artigos seguiram a recomendação do periódico analisado.

Quadro 1 - Palavras-chave X Descritores

Artigo	Palavras-chaves	Descritores DeCS
1	AIDS; Direitos Sexuais e Reprodutivos; Gênero e Saúde; Mulheres	Direitos Sexuais e Reprodutivos; Gênero e Saúde; Mulheres
2	Avaliação em Saúde; Mortalidade Neonatal Precoce; Near Miss; Recém-Nascido; Sistemas de Informação	Avaliação em Saúde; Mortalidade Neonatal Precoce; Near Miss; Recém-Nascido; Sistemas de Informação
3	Doença de Chagas; Técnicas de Apoio para a Decisão; Vigilância em Saúde Pública	Doença de Chagas; Técnicas de Apoio para a Decisão; Vigilância em Saúde Pública
4	Análise Espaço-Temporal Coronavirus; Geografia Médica;	Análise Espaço-Temporal Coronavirus; Geografia Médica
5	Aborto; Aborto Induzido; Direitos Sexuais e Reprodutivos; Mortalidade Materna	Aborto; Aborto Induzido; Direitos Sexuais e Reprodutivos; Mortalidade Materna
6	Anticoncepção; Atenção Primária à Saúde; Avaliação de Serviços de Saúde; Estudos Transversais; Sistema Único de Saúde	Anticoncepção; Atenção Primária à Saúde; Estudos Transversais; Sistema Único de Saúde
7	Diálise; Insuficiência Renal Crônica; Transplante de Rim	Diálise; Insuficiência Renal Crônica; Transplante de Rim
8	Atenção Primária à Saúde; Difusão de Inovação; Sistemas de Informação em Saúde; Utilização de Equipamentos e Suprimentos	Atenção Primária à Saúde; Difusão de Inovação; Sistemas de Informação em Saúde; Utilização de Equipamentos e Suprimentos

Fonte: elaborado pelas autoras.

A partir da mineração de textos realizada via software Sobek, apresenta-se a análise dos conteúdos dos artigos.

Observa-se que o artigo 1 apresenta uma grande coerência em sua representação, isto é, os termos de maior frequência do texto foram considerados em suas palavras-chave (quadros 1 e 2). Porém, nota-se que o termo HIV não foi utilizado como palavra-chave, mesmo sendo um conceito reconhecido no texto, possuindo diversas conexões (arestas) com outros conceitos identificados e um descritor autorizado pelo DeCS (quadro 2).

Quadro 2 - Conceitos localizados no artigo 1

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Saúde	75	HIV; HIV Aids; Gênero; Direitos Sexuais; Sexual
HIV	31	Saúde
Mulher	30	Sexual
Vida	26	Sexual; Social
Sexual	26	Saúde; Mulher; Vida
Direitos Sexuais	24	Saúde
Gênero	24	Saúde
HIV Aids	24	Saúde
Social	24	Vida

Fonte: elaborado pelas autoras.

No artigo 2 a coerência também é grande entre as palavras-chave e o conteúdo do texto (quadros 1 e 3). Mas os termos Avaliação em Saúde e Sistemas de Informação, mesmo sendo palavras-chave e descritores no artigo, não se apresentam na lista de termos mais representativos do texto, como visto no quadro 3.

Quadro 3 - Conceitos localizados no artigo 2

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Neonatal	155	Near Miss
Near Miss	100	Neonatal
Vivos	51	Nascidos
Casos	45	----------
Recém-Nascidos	45	----------
Nascidos	42	Vivos
Brasil	40	----------
Mortalidade	36	----------
Saúde Pública	35	----------
IMIP (Instituto de Medicina Integral Professor Fernando Figueira)	34	HC
HC (Hospital das Clínicas)	28	IMIP
Nascimento	27	----------

Fonte: elaborado pelas autoras.

No artigo 3 existe uma coerência média entre os conceitos mais frequentes do texto e suas palavras-chave (quadros 1 e 4). A palavra-chave Técnicas de Apoio para a Decisão,

mesmo sendo um descritor autorizado não está na lista de termos mais representativos do texto, como visto no quadro 4.

Quadro 4 - Conceitos localizados no artigo 3

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Chagas	96	Crônica; Doença
Doença	83	Chagas; Crônica
Municípios	53	----------
Crônica	31	Chagas; Doença
Indicadores	27	----------
Análise	20	Multicritério
Casos	20	----------
População	20	----------
Saúde Pública	20	----------

Fonte: elaborado pelas autoras.

O artigo 4 apresenta baixa coerência entre as palavras-chave e seu conteúdo (quadros 1 e 5). O que ocorre nesse artigo parece ter um detalhe específico: por ser tratar de um tema emergente, seu termo mais representativo - COVID-19 - no momento de sua elaboração poderia ainda não ser um termo autorizado pelo DeCS, o que pode ter influenciado na escolha das palavras-chave.

Quadro 5 - Conceitos localizados no artigo 4

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
COVID	74	Santa Catarina; Casos; Dados; Trabalho; Estado; Catarinense; Clusters; Scan
Catarinense	54	Santa Catarina; COVID; Dados
Santa Catarina	38	COVID; Estado; Catarinense; Dados
Casos	33	COVID; Dados; Catarinense
Dados	31	COVID; Santa Catarina; Dados
Clusters	24	COVID; Scan
Trabalho	24	COVID
Estado	23	Santa Catarina; COVID
Scan	23	Clusters; COVID

Fonte: elaborado pelas autoras.

No artigo 5, os conceitos mais representativos do texto apresentam grande coerência com suas palavras-chave (quadros 1 e 6).

Quadro 6 - Conceitos localizados no artigo (5)

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Aborto	120	Ministério da Saúde; Mulheres; Óbitos
Óbitos	44	Aborto
Rio de Janeiro	38	Brasil
Mortalidade	37	----------

Mulheres	37	Aborto
Ministério da Saúde	34	Aborto
Brasil	33	Rio de Janeiro
Comissões	33	Aborto

Fonte: elaborado pelas autoras.

No artigo 6, os termos mais representativos do textos não se apresentam nas palavras-chave, mostrando baixa, ou nenhuma, coerência entre seu conteúdo e seus termos representativos (quadros 1 e 7).

Quadro 7 - Conceitos localizados no artigo (6)

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Disponibilidade	67	Ciclos; Insumos; UBS
UBS (Unidade Básica de Saúde)	57	Ciclos; Disponibilidade
Insumos	42	Disponibilidade;Planejamento
Ciclos	41	PMAQ-AB; Disponibilidade; UBS
PMAQ-AB (Programa Nacional de Melhoria do Acesso e da Qualidade da Atenção Básica)	40	Ciclos; UBS
Cobertura	30	ESF; IDH
ESF (Estratégia Saúde da Família)	30	Cobertura; Porte
Planejamento	29	Insumos
IDH (Índice de Desenvolvimento Humano)	28	Cobertura
Porte	28	ESF

Fonte: elaborado pelas autoras.

Já no artigo 7 ocorre média coerência entre os conceitos mais representativos do texto e as palavras-chave (quadros 1 e 8).

Quadro 8 - Conceitos localizados no artigo (7)

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Transplante Renal	110	Acesso; Brasil; Pacientes
Pacientes	91	Transplante Renal
Acesso	41	Transplante Renal
DRC (Doença Renal Dialítica)	40	----------
Estudo	31	----------
Brasil	26	Transplante Renal

Fonte: elaborado pelas autoras.

O artigo 8 apresenta média coerência entre os conceitos mais frequentes de seu conteúdo e suas palavras-chave (quadros 1 e 9). Mas nota-se que o termo Sistema Único de Saúde (SUS) não foi utilizado como palavra-chave: mesmo sendo um conceito de alta frequência

no texto, com diversas conexões (arestas) com os outros conceitos identificados e um descritor autorizado pelo DeCS (quadro 4).

Quadro 9 - Conceitos localizados no artigo (8)

Conceito identificado/Sobek	Número de vezes que o conceito aparece no texto	Arestas do conceito
Uso	72	Atributos; Inovação
Estudo	53	Dados; SUS
Inovação	53	Adoção; Uso; SUS; APS
Sistema	44	Brasil
Atributos	43	Uso
Adoção	35	Inovação
Dados	33	Estudo; SUS
SUS	33	Estudo; Inovação
APS (Atenção Primária à Saúde)	29	Inovação
Resultados	27	----------
Brasil	25	Sistema

Fonte: elaborado pelas autoras.

5 CONSIDERAÇÕES FINAIS

Observou-se que uma grande coerência semântica aconteceu somente nos artigos 1, 2 e 5. Nos outros artigos estudados, o nível de coerência ficou entre baixo e médio: em alguns casos os termos de alta frequência não foram utilizados, e em outros casos as palavras-chave foram utilizadas mesmo não estando na lista dos termos mais frequentes do artigo. Esse resultado pode refletir uma escolha consciente dessas palavras-chave pelos autores, ou o desconhecimento do vocabulário controlado da área e de seus recursos.

É possível considerar como a obrigatoriedade do uso de um vocabulário influencia na representação temática de um artigo. Como foi escrito no início deste trabalho, os artigos estudados foram publicados num periódico que recomenda aos autores que as palavras-chave apresentadas no artigo devem estar no vocabulário da área. Reconhece-se a importância do uso da terminologia em áreas específicas, mas sua obrigatoriedade pode interferir na atualização e inovação dessas áreas.

A partir deste exercício exploratório, pode-se considerar que a mineração de texto e a indexação automática podem colaborar com a coerência semântica na representação do conteúdo de artigos científicos. Embora os sintagmas nominais também sejam considerados como formas de representação desse conteúdo - agregando valor semântico à descrição do artigo, conforme apontados pelos autores Kuramoto (1995; 2002) e Correa e Celerino (2019) -

esse tipo de análise não foi realizada neste estudo, mas poderá ser objeto de aprofundamento futuramente.

Compreendendo a indexação como representação de conteúdo e seu importante papel no sistema de comunicação na ciência, procurou-se refletir sobre a informação em saúde representada nesses textos, entendendo a relevância da visibilidade e recuperação dessa produção. Vale ressaltar que o periódico científico - e por conseguinte seus artigos - além de outros papéis estratégicos na comunicação da ciência, é registro da memória, chancela de qualidade e difusor do conhecimento (GUANAES; GUIMARÃES, 2012).

Por fim, cabe informar que este trabalho faz parte de um estudo exploratório para uma tese de doutoramento já qualificada, e que serão feitas experiências com outros softwares que poderão ser utilizados por profissionais de informação envolvidos na representação temática em ambientes digitais.

REFERÊNCIAS

ARAÚJO JÚNIOR, Rogério H. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007.

BOCCATO, Vera R. Casari; TORQUETTI, Melissa Camargo. Interoperabilidade entre linguagens de indexação como recurso de modelagem de repertório terminológico de coordenadorias de comunicação social em ambientes universitários: uma proposta metodológica. **Informação & Informação**, v. 17, n. 3, p. 76-101, 2012.

CURTY, Marlene Gonçalves; BOCCATO, Vera R. Casari. O artigo científico como forma de comunicação do conhecimento na área de Ciência da Informação. **Perspectiva em Ciência da Informação**, v.10, n.1, p.94-107, 2005. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/305>.

CORRÊA, Renato Fernandes; CELERINO, Victor Galvão. Método de normalização de sintagmas nominais na indexação automática. **Em Questão**, v.25, n.1, p.321-344, 2019. Disponível em: <https://doi.org/10.19132/1808-5245251.321-344>.

DeCS: Descritores em Ciências da Saúde. Disponível em: <https://decs.bvsalud.org/>. Acesso em: 20 fev. 2021.

DIAS, Eduardo Wense; NAVES, Madalena M. Lopes. **Análise de assunto**: teoria e prática. 2.ed. Brasília: Briquet de Lemos, 2013.

FÓRIS, Ágota. Network theory and terminology. **Knowledge Organization**, v.40, n.6, p.422-429, 2013.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO (FAPESP). REVISTA PESQUISA (FAPESP). Publicações científicas por países: contagem por autoria e por artigo. **Rev. Pesquisa FAPESP**, n.288, 2020. Disponível em: <https://revistapesquisa.fapesp.br/publicacoes-cientificas-por-paises-contagem-por-autoria-e-por-artigo/>.

GIL LEIVA, Isidoro. 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.

GUANAES, Paulo C. Vieira; GUIMARÃES, Maria Cristina S. Modelos de gestão de revistas científicas: uma discussão necessária. **Perspectivas em Ciência da Informação**, v.17, n.1, p.56-73, 2012. Disponível em: <https://doi.org/10.1590/S1413-99362012000100004>.

KRIEGER, Maria da Graça; FINATTO, Maria José B. **Introdução à terminologia**: teoria e prática. São Paulo: Contexto, 2004.

KURAMOTO, Helio. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero,** v.3, n.1, 9p., 2002.

KURAMOTO, Helio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, v.25, n.2, p.1-18, 1995.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática 2. ed. Brasília,DF: Briquet de Lemos/Livros, 2004.

LAPA, Remi C; CORRÊA, Renato F. Indexação automática no âmbito da ciência da informação no Brasil. **Informação & Tecnologia**, v.1, n.2, p.59-76, 2014.

LATOUR, Bruno; WOOLGAR, Steve. A credibilidade cientifica. In: **A vida de laboratório**. Rio de Janeiro: Relume Dumará, 1997.

LIDDY, Elizabeth D. **Natural language processing.** Disponível em: <https://surface.syr.edu/cgi/viewcontent.cgi?article=1019&context=cnlp>. Acesso em: 10 mar. 2021.

MEADOWS, Arthur J. Canais da comunicação científica. In: MEADOWS, Arthur J. **A comunicação científica**. Brasília,DF: Briquet de Lemos, 1999.

MOTA, Daniel M; FERREIRA, Paulo J. G; LEAL, Lisiane F. Produção científica sobre a COVID-19 no Brasil: uma revisão de escopo. **Visa em Debate**, v.8, n.3, p.114-124, 2020. Disponível em: <https://doi.org/10.22239/2317-269x.01599>.

OLIVEIRA, Lais Pereira et al. Política de indexação em periódicos da Ciência da Informação: um estudo das diretrizes para atribuição de palavras-chave aos artigos. **Perspectivas em Ciência da Informação**, v.25, n.4, p.140-169, 2020. Disponível em: <https://doi.org/10.1590/1981-5344/3876>.

PLATAFORMA SUCUPIRA. **Qualis periódicos:** classificação de periódicos quadriênio 2013-2016. Disponível em:

<https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 11 jun. 2022.

PUERTA-DÍAZ, Mirelys et al. O processamento de linguagem natural nos estudos métricos da informação: uma análise dos artigos indexados pela Web of Science (2000-2019). **Encontros Bibli: Rev. eletrônica de Biblioteconomia e Ciência da Informação,** v.26, p.1-24, 2021. Disponível em: <https://doi.org/10.5007/1518-2924.2021.e76886>.

SALES, Rodrigo de. Suportes teóricos para pensar linguagens documentárias. **Rev. digital de Biblioteconomia e Ciência da Informação**, v.5, n.1, p.96-114, 2007.

SANTIN, Dirce M.; CAREGNATO, Sônia E. Produção e impacto científico da América Latina e Caribe em revistas: um olhar sobre a ciência local e global. In: CARNEIRO, Felipe F. Barros; FERREIRA NETO, Amarílio; SANTOS, Wagner. **A comunicação cientifica em periódicos**. Curitiba: Appris, 2019.

SciELO SAÚDE PÚBLICA. Disponível em: <https://www.scielo20.org/redescielo/wp-content/uploads/sites/2/2018/09/Informe-SciELO-Sa%C3%BAde-P%C3%BAblica.pdf>. Acesso em: 1 mar. 2021.

TONELLO, Izangela M.S; LUNARDELLI, Rosane S. Alvares; ALMEIDA JUNIOR, Oswaldo Francisco. Palavras-chave: possibilidades de mediação da informação. **Ponto de Acesso**, v.6, n.2, p.21-34, 2012.

UNIVERSIDADE DE BRASÍLIA (UNB). **Prévia do Qualis CAPES periódico quadriênio 2017-2020 (provisório).** Disponível em: <https://cen.unb.br/posgrad/documentos/item/358-previa-do-qualis-capes-periodico-quadrienio-2017-2020-provisorio>. Acesso em: 11 jun 2022.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS). PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA NA EDUCAÇÃO (PPGIE). **Sobek mining**. Disponível em: <http://sobek.ufrgs.br>. Acesso em: 13 maio 2021.