

ATRIBUIÇÃO DE PALAVRAS-CHAVE EM TRABALHOS COMPLETOS DE EVENTO: ANÁLISE DE CONTROLE DE VOCABULÁRIO

ASSIGNMENT OF KEYWORDS FOR COMPLETE WORKS AT THE EVENT: ANALYSIS OF VOCABULARY CONTROL

Mariangela Spotti Lopes Fujita^a
Richele Grenghe Vignoli^b
Eder Antonio Pansani Junior^c
Paula Regina Dal'Evedove^d

RESUMO

Objetivo: Com o objetivo de verificar o processo de representação na atribuição de palavras-chave em publicações de evento em Ciência da Informação, foi realizada investigação sobre o controle de vocabulário com uso do Tesouro Brasileiro em Ciência da Informação (TBCI) pelos autores no Portal do Encontro da Associação Nacional de Pesquisa em Ciência da Informação (ENANCIB). **Metodologia:** O estudo teve como universo de pesquisa os Anais do ENANCIB e, como amostra, foram analisadas publicações da modalidade de trabalhos completos dos Anais de 2015, ano em que o controle do vocabulário começou a ser recomendado até 2018. Realizou-se análise das palavras-chave atribuídas pelos autores, após extração e mineração com o *Software* de Extração de Palavras-Chave (SEP), criado para essa finalidade. **Resultados:** Foram recuperados 1.075 artigos dos 11 Grupos de Trabalho do evento com um total de 4.437 palavras-chave examinadas. Os principais resultados demonstram que 60,3% das palavras-chave recuperadas nos artigos estão presentes como termos controlados no Tesouro. O Grupo de Trabalho 4 obteve o melhor desempenho, na análise efetuada, e o 10, o resultado mais incipiente. **Conclusões:** Os dados coletados indicam que mudanças são necessárias para que os Anais do ENANCIB possuam um controle de vocabulário mais abrangente. Recomenda-se a elaboração de tutoriais aos autores, com orientações e benefícios sobre o uso do Tesouro para controle de vocabulário.

^a Doutora em Ciências da Comunicação pela Universidade de São Paulo (USP). Docente do Programa em Ciência da Informação da Universidade Estadual Paulista (UNESP) e Assessora e Presidente do Comitê de área de Comunicação, Artes e Ciência da Informação do CNPq, Marília, Brasil. E-mail: mariangela.fujita@unesp.br

^b Doutora em Ciência da Informação pela Universidade Estadual Paulista (UNESP), Marília, Brasil. E-mail: rivignoli@gmail.com

^c Doutor em Ciência da Informação pela Universidade Estadual Paulista (UNESP). Docente do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Votuporanga, Brasil. E-mail: epansani@gmail.com

^d Doutora em Ciência da Informação pela Universidade Estadual Paulista (UNESP). Docente do Departamento de Ciência da Informação na Universidade Federal de São Carlos (UFSCAR), São Carlos, Brasil. E-mail: dalevedove@ufscar.br

Descritores: Indexação por autores. Controle de vocabulário. Tesouro. Palavras-chave.

1 INTRODUÇÃO

O autor é um especialista de domínio, o qual, quando realiza a atribuição de palavras-chave, o faz com o objetivo de que o conteúdo de seu texto seja recuperado e lido por outros, igualmente especialistas de domínio. Isso faz com que essas palavras-chave tenham relevância para a ampliação e a inovação de vocabulários especializados, tendo em vista a necessidade de atualização constante. Sob o ponto de vista de sua área de conhecimento, é o melhor indexador e, desse modo, as palavras-chave que atribui são relevantes para os especialistas do domínio de conhecimento. Contudo, em contexto mais generalizado, tais palavras-chave podem nunca ser recuperadas em uma busca, a não ser por outros especialistas no domínio.

Por outro lado, os objetivos do indexador profissional vão além da representação por palavras-chave, que serão recuperadas por outros especialistas de domínio e, por isso, adota vocabulários controlados, considerando a necessidade de muitos usuários que podem não ser os especialistas de domínio. Esse objetivo é pertinente também com respeito aos autores, porque buscam periódicos e fontes indexadores que sejam amplamente acessadas por outros autores, os quais possam citar sua publicação.

A principal vantagem das palavras-chave para os editores de publicações científicas está na economia de tempo, em relação à indexação por profissionais. Cada autor pode atribuir palavras-chave sem que haja custo operacional, financeiro ou temporal. Hartley e Kostoff (2003) elencam os principais benefícios das palavras-chave, em uma publicação especializada:

1. permitem que os leitores decidam se um artigo possui ou não material relevante de acordo com seus interesses;
2. fornecem aos leitores termos para uso em buscas de documentos similares na *web*;
3. auxiliam indexadores e editores a agrupar materiais relacionados;
4. possibilitam que editores e pesquisadores registrem mudanças nos assuntos de uma disciplina, através do tempo;

5. ligam tópicos específicos de pesquisa, dentro de um escopo maior.

Essas vantagens são oportunamente assumidas, sobretudo, por mecanismos de buscas em que a pesquisa por palavras-chave se tornou o método mais usado e preferido por qualquer sistema *on-line*, mas as opções de pesquisa avançada ajudam os usuários finais a refinar pesquisas, navegar e gerenciar grandes conjuntos de resultados (GROSS; TAYLOR; JOUDREY, 2015).

A desvantagem é justamente a falta do método de indexação que melhor realiza a representação do conteúdo indexado com vocabulário controlado. Contudo, conforme Schwing, Mccutcheon e Maurer (2012, p. 923-924), sobre a atribuição de palavras-chave por autores de teses e dissertações, “[...] os pontos fracos das palavras-chave são os pontos fortes de um vocabulário controlado e vice-versa [...]”, pois as palavras-chave representam ideias e conceitos atuais das pesquisas feitas por publicações, teses e dissertações e trabalhos completos de eventos, e os termos do vocabulário controlado tendem a ser mais estáveis e ter relações com assuntos mais amplos.

Apesar da larga utilização que a atribuição de palavras-chave tem, nos sistemas de autoarquivamento, em repositórios digitais, periódicos e eventos científicos, nem os autores e nem os sistemas parecem ter consciência dos benefícios que o processo de indexação com controle de vocabulário proporciona. Pesquisas recentes dedicadas à questão evidenciam a falta de orientação metodológica sobre o processo de indexação, variáveis e ferramentas de controle de vocabulário em sistemas de autoarquivamento, tais como um tesouro à disposição dos autores, quando da submissão de suas publicações (FUJITA; AGUSTÍN-LACRUZ; TERRA, 2018; FUJITA; TARTAROTTI, 2020; TERRA; AGUSTÍN-LACRUZ; BERNARDES; FUJITA, 2021). Nesse sentido, é importante efetuar estudos sobre o processo de representação por palavras-chaves, em publicações e trabalhos acadêmicos, na perspectiva da indexação promovida pelo autor pesquisador, compreendendo tal sujeito como um indexador especialista, o qual domina a terminologia de um determinado campo científico (COSTA; MOURA, 2013).

Com o objetivo de verificar o processo de representação na atribuição de palavras-chave em publicações de evento em Ciência da Informação, foi

realizada investigação sobre o controle de vocabulário com uso do Tesouro Brasileiro em Ciência da Informação (TBCI) pelos autores no Portal do Encontro da Associação Nacional de Pesquisa em Ciência da Informação (ENANCIB), mediante análise das diretrizes aos autores e das palavras-chave atribuídas pelos autores pesquisadores, após extração e mineração com o *Software* de Extração de Palavras-Chave (SEP), criado para essa finalidade.

O *software* SEP foi desenvolvido pelos autores, empregando tecnologias e linguagens *web*, como PHP (acrônimo recursivo, *PHP: Hypertext Preprocessor*), HTML (*HyperText Markup Language*), CSS (*Cascading Style Sheets*) e *Framework Bootstrap*¹, a fim de criar uma interface amigável pela qual foi possível extrair as palavras-chave de cada artigo. A aplicação foi projetada especificamente para essa finalidade, assim como no tratamento dos dados extraídos para apresentação de relatórios usados para as análises efetuadas no contexto deste estudo.

2 O TESAURO NO CONTROLE DE VOCABULÁRIO

O multiculturalismo e influências midiáticas fazem com que a linguagem, essencialmente nas sociedades modernas, esteja em constante movimento e renovação. Nas mudanças linguísticas e terminológicas, domínios do conhecimento são incitados a uma revisão e atualização, a qual possibilite o desenvolvimento de seu arcabouço teórico e prático. Da mesma forma, vocabulários controlados necessitam ser revisados e atualizados, com o intuito de favorecer a plena organização e recuperação da informação de seus domínios.

No contexto da *web*, os vocabulários controlados se tornam cada vez mais necessários, de sorte a individualizar e padronizar termos de um domínio do conhecimento. Dentre os vocabulários controlados mais conhecidos estão os glossários, tesouros, sistemas de classificação, listas de cabeçalhos de assunto, taxonomias, ontologias, anéis de sinônimos, entre outros.

Ora, os vocabulários controlados, conforme Lopes (2002), são conjuntos de termos organizados hierarquicamente ou alfabeticamente, os quais objetivam a

¹ Disponível em: <https://getbootstrap.com/>. Acesso em: 10 jan. 2023.

recuperação da informação por temas, além de reduzir a diversidade terminológica.

Devido à sua importância, “[...] o controle de vocabulário é um recurso para organizar e recuperar [...] informações com consistência, gerando, conseqüentemente, confiança no sistema.” (SMIT; KOBASHI, 2003, p. 14). A confiança nos sistemas de informação relatados pelas autoras pode ser compreendida como a sua capacidade de recuperar a informação precisamente, de acordo com as necessidades de seus usuários. Para tanto, um controle de vocabulário necessita ser realizado, para que a busca por informação aconteça de forma eficaz e para que o vocabulário controlado corresponda às expectativas de seu funcionamento, juntamente ao domínio e a seu público.

Smit e Kobashi (2003) demonstram outro olhar para definir as funções do controle de vocabulário e vocabulário controlado, em sistemas de informação:

- Controle de Vocabulário: processo, um objetivo que se deseja atingir;
- Vocabulário Controlado: instrumento (ou produto) para nomear atividades/funções (documentos) que geram confiança no sistema.

A partir do ponto de vista de Smit e Kobashi (2003), o controle de vocabulário pode ser entendido como o processo ou a ação de controlar a escolha dos termos usados naquele sistema. O controle é imprescindível, para que unidades preferenciais sejam interligadas a conceitos ou noções, em linguagens documentárias. O controle do vocabulário de um sistema de informação deve ser efetuado com vistas a padronizar os termos utilizados naquele sistema, seja por profissional da informação, seja por usuários/sujeitos de informação.

Um vocabulário controlado pode ser confeccionado segundo as peculiaridades de cada instituição ou baseado em outros instrumentos existentes, que possuem características em comum com o público e/ou domínio ao qual se destinará. A representação documentária está presente tanto nos vocabulários controlados, por meio dos termos antes traduzidos por linguagens de indexação, quanto no controle do vocabulário, o qual deve se fundamentar em padronização dos termos empregados em cada demanda informacional.

Para Wu (2018), o contexto de informações na *web*, assim como o desenvolvimento de sistemas e bibliotecas digitais, fizeram com que os tesouros voltassem a atrair a atenção de estudos, na Ciência da Informação. O foco,

então, se volta para as formas pelas quais o instrumento deve ser aplicado na recuperação conceitual da informação na *web*, na gestão do conhecimento e da informação e como suporte para os estudos de ontologias e ciência de dados. No entanto, na edição da *International Society for Knowledge Organization* (ISKO-UK 2015), o papel do tesouro na recuperação de informação moderna foi tomado como inferior ao que já fora antes (HJØRLAND, 2016). Ao questionar as projeções dos tesouros na *Library and Information Science* (LIS), Hjørland (2016, p. 145, tradução nossa) declara que o “[...] tesouro tradicional não tem lugar na recuperação de informação moderna, mas que ferramentas semânticas mais flexíveis baseadas em estudos próprios de domínios serão sempre importantes.”

Todavia, os tesouros tradicionais, aqueles tidos como estáticos e sem nenhuma atualização ou acréscimo de termos, são desafiados ao descaso e desuso, frente a certos buscadores modernos, como o *Google* (HJØRLAND, 2016; WU, 2018). No raciocínio de Hjørland (2016), a Ciência da Informação deverá fortalecer os estudos dos tesouros, com forte apoio na terminologia e no conhecimento compreendido com relevância em diversos domínios. Em outras palavras, é preciso estabelecer relações semânticas e conceituais com outras disciplinas, a fim de que a recuperação da informação se concretize, através de tesouros modernos, especialmente em ambiente *web*. É esperado, do mesmo modo, que o conhecimento organizado para um domínio seja útil para outros.

A revisão ou atualização de tesouros como partícipe da agenda de pesquisa da Ciência da Informação e Organização do Conhecimento implica, entre outros procedimentos, a inserção, modificação e exclusão de termos e registros em bases de dados indexadas (FERREIRA; MACULAN, 2020), além de adequação às Tecnologias da Informação e Comunicação (TIC). De acordo com Ferreira e Maculan (2020), a implementação de revisão e atualização de tesouros justapõe-se à adoção de diretrizes científicas e normativas que possibilitem a modelização e/ou reestruturação conceitual dos instrumentos terminológicos.

A revisão e a atualização de tesouros constituem uma ação complexa, a qual exige esforços múltiplos de profissionais especializados, além de uso de aportes conceituais e normativos que validem os procedimentos. Ferreira e Maculan (2020) aludem a alguns instrumentos que possibilitam a reestruturação conceitual e

terminológica de tesouros, que se estendem ao desenvolvimento de categorias, classes, subclasses, definições e relações semânticas, como, por exemplo, a Teoria da Classificação Facetada (TCF) de Shiyali Ramamrita Ranganathan; a Teoria do Conceito de Ingetraud Dahlberg; a Teoria Linguística de John Lyons; a Teoria Terminológica de Eugen Wüster; as Teorias Semânticas de Bernard Pottier, Charles J. Fillmore e Jean Claude Gardin, e a Norma ISO 25.964-1 (INTERNATIONAL STANDARD ORGANIZATION, 2011), a qual, segundo Hjørland (2016), é importante para os profissionais da informação, mas não é suficiente para elevar os tesouros a um estado de atualização e modernização.

Além das teorias, são necessários outros aportes e conhecimentos da Ciência da Informação e Organização do Conhecimento, como o insumo à elaboração de definições, revisões e possíveis atualizações aos tesouros. Hjørland (2016) sugere que o *design* dos tesouros deveria ser revisto e que sua avaliação fosse feita por atualizações com o emprego de ferramentas semânticas mais flexíveis, como os *Topic Maps*, ontologias e uso de linguagens filosóficas. A defesa do autor é por instrumentos semânticos flexíveis e não por padrões que não permitam ou dificultem a atualização constante dos tesouros, como são os tesouros tradicionais (HJØRLAND, 2016).

Em busca de demonstrar o dinamismo da linguagem e a necessidade de atualização dos tesouros, Wu (2018) elaborou uma pesquisa para explicar a respeito do acréscimo de termos relacionados, com base na percepção de usuários. A ideia foi exemplificar que, se um tesouro tradicional for enriquecido, possivelmente apresentará melhora na organização e na descoberta de conhecimentos. Com o auxílio de um dicionário de sinônimos, a investigação de Wu (2018) constata que, se um tesouro cobre os conceitos principais de um domínio, um tesouro enriquecido é útil para que os leitores possam compreender melhor o conhecimento do domínio, além de ensejar novas descobertas.

Estudos sobre uso de palavras-chave atribuídas por autores são importantes para a avaliação e atualização de vocabulários controlados. Strader (2009) alerta, porém, para o fato de que a maioria dos estudos sobre assunto se concentram em usuários como buscadores de informações e menos no uso de palavras-chave atribuídas por autores de documentos online. Em sua pesquisa,

Strader (2009) investigou justamente sobre a sobreposição entre palavras-chave atribuídas pelo autor e os cabeçalhos de assunto da Biblioteca do Congresso (LCSH) atribuídos pelo catalogador em teses e dissertações no catálogo online da Ohio State University com a intenção de contribuir com a literatura sobre o debate entre palavras-chave versus vocabulários controlados em catálogos online e palavras-chave.

Entretanto, estudos de avaliação de palavras-chaves realizados por, Moraes e Campos (2014), Martins, Mota e Felipe (2017) e Almeida (2019) investigaram as palavras-chaves de trabalhos apresentados no Grupo de Trabalho Organização e Representação da Informação (GT2) do Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB para determinação de classes temáticas e também para análise da representação por palavras-chave dos temas de produções científicas dos Programas de Pós-Graduação em Ciência da Informação.

Diferentes metodologias foram aplicadas pelos autores para o mapeamento e análise das palavras-chaves. Moraes e Campos (2014) analisaram 275 palavras-chave das comunicações orais do Grupo de Trabalho 2 - Organização e Representação da Informação (GT2) do Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB no período de 2003 a 2013 realizando um mapeamento temático com uso do CKSOL para classificação dos termos. Martins, Mota e Felipe (2017), avaliaram 975 palavras-chave dos trabalhos apresentados no período de 2003 a 2016 de acordo com as 10 classes temáticas do CSKOL. Almeida (2019) realizou um mapeamento das classes do CSKOL aplicado à artigos e ementas de disciplinas com a finalidade de identificar sua comunidade discursiva no domínio da área de Organização do Conhecimento. Os trabalhos citados são coincidentes na metodologia de análise por palavras-chave com resultados utilizados para avaliação de domínio de conhecimento especializado o que revela a relevância temática das palavras-chaves na representação de assuntos, especialmente, no que se refere à área de Organização e Representação do Conhecimento.

Considerando que o Tesouro Brasileiro de Ciência da Informação passou a ser indicado nas orientações aos autores do ENANCIB para controle de vocabulário importa conhecer as orientações aos autores sobre o uso do

vocabulário controlado e sua função de controle de vocabulário, assim como avaliar o uso do Tesouro por meio da análise das palavras-chave. Compreendendo que o uso primário de vocabulários controlados do tipo tesouro é viabilizar a representação documentária para conferir qualidade e padronização das publicações de um determinado campo científico, no apoio ao processo de busca e recuperação, de maneira mais precisa, o emprego efetivo de tesouros, por parte dos pesquisadores, é uma prática desejável para a atribuição de palavras-chave com controle de vocabulário. Nesse processo, atualizações nos vocabulários controlados, em atenção às expansões e transformações das áreas de conhecimento para os quais foram elaborados, contribuem para que pesquisadores recorram ao instrumento, no processo de escolha e atribuição de palavras-chave de suas pesquisas. Por conseguinte, o controle de vocabulário está intimamente associado à disponibilização de instrumentos atualizados e representativos, os quais ofereçam termos compatíveis com as temáticas e conteúdos tratados pelos pesquisadores.

3 METODOLOGIA

O estudo teve como universo de pesquisa os Anais do ENANCIB, um dos eventos mais conceituados da Ciência da Informação, no Brasil; como amostra, foram analisadas publicações da modalidade de comunicação oral provenientes dos Anais de 2015, ano em que o controle do vocabulário começou a ser recomendado no portal do evento, até 2018. O escopo do estudo compreendeu pesquisas publicadas na modalidade de trabalhos completos (escritos por e entre doutores e/ou com parceria de doutorandos e mestres). Portanto, os trabalhos que compuseram a amostra desta pesquisa foram apresentados durante o evento e publicados em formato completo, nos Anais do ENANCIB, nos anos de 2015 a 2018.

Os procedimentos de análise do controle de vocabulário das palavras-chave atribuídas pelos autores de trabalhos completos, em metadados de assuntos, durante o autoarquivamento no sistema de evento científico, foram executados em duas etapas: pela verificação das diretrizes aos autores, nos anos da amostra, no Portal do evento; e pela extração e mineração das palavras-

chave contidas nos artigos dos 11 Grupos de Trabalhos (GT),² por meio da criação do *Software* de Extração de Palavras-Chave (SEP).

O SEP foi criado com base em linguagem de programação PHP³ e banco de dados *MySQL*. Seu funcionamento consistiu na conversão dos arquivos em PDF para textos sem formatação. Na sequência, foi feita uma análise nos arquivos, em busca do rótulo *Palavras-chave*. Assim, 4.437 palavras-chave foram mineradas e armazenadas no banco de dados do SEP. O programa ainda registrou informações importantes, como título, ano e GT de cada um dos trabalhos completos analisados. Na Figura 1 pode-se observar uma das telas do SEP, responsável por permitir a conferência e o ajuste das palavras-chave, via processamento automático:

Figura 1 – Tela de edição do SEP

Formulário de edição de documentos

Filename: 150-3549-1-PB.pdf

Edição: XVIII ENANCIB Ano: 2017

GT: GT-1 - Estudos Históricos e Epistemológicos da Ciência da Informação

Tipo: Comunicação Oral

Título: MANUEL CÍCERO PEREGRINO DA SILVA NA BIBLIOTECA NACIONAL: ENGAJAMENTO AOS IDEIAS DE OTLET E LA FONTAINE E A DOCUMENTAÇÃO

Enviar Limpar

Palavra Chave 1: Biblioteca Nacional

Palavra Chave 2: Documentação

Palavra Chave 3: Instituto Internacional de Bibliografia

Palavra Chave 4: Manoel Cícero Peregrino da Silva

Palavra Chave 5: Paul Otlet

Palavra Chave 6:

Palavra Chave 7:

Palavra Chave 8:

XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-1 - Estudos Históricos e Epistemológicos da Ciência da Informação

MANUEL CÍCERO PEREGRINO DA SILVA NA BIBLIOTECA NACIONAL: ENGAJAMENTO AOS IDEIAS DE OTLET E LA FONTAINE E A DOCUMENTAÇÃO

Carlos Henrique Juvêncio (Universidade de Brasília – UnB)

Georgete Medlag Rodrigues (Universidade de Brasília – UnB)

MANUEL CÍCERO PEREGRINO DA SILVA IN THE NATIONAL LIBRARY: ENGAJEMENT TO THE IDEALS OF OTLET AND LA FONTAINE AND DOCUMENTATION

Modalidade da Apresentação: Comunicação Oral

Resumo: Os belgas Paul Otlet e Henri La Fontaine, em finais do século XIX, criam o Instituto Internacional de Bibliografia que, para além de criar uma grande fonte de informação universal, foi pioneiro na adoção de novos modos de tratamento dos acervos documentais, inaugurando a concepção de Documentação. Manoel Cícero Peregrino da Silva, diretor da Biblioteca Nacional brasileira entre 1900 e 1924, adotou várias mudanças na instituição, como a inauguração de um novo prédio, reorganização de seu acervo, criação de novos serviços e, em certa medida, segue a trilha aberta por Otlet e La Fontaine. Essa comunicação faz um inventário das ações de Peregrino da Silva buscando entender em que medida essas ações contribuíram para a Documentação no Brasil e seu

Fonte: Elaborada pelos autores.

Outra função do SEP foi gerar relatórios, apresentações dos dados coletados em formatos úteis aos propósitos do estudo, assim como a apuração das palavras-chave mais recorrentes por GT, no período da amostra. Apesar da

² GT 1 Estudos Históricos e Epistemológicos da Ciência da Informação; GT 2 Organização e Representação do Conhecimento; GT 3 Mediação, Circulação e Apropriação da Informação; GT 4 Gestão da Informação e do Conhecimento; GT 5 Política e Economia da Informação; GT 6 Informação, Educação e Trabalho; GT 7 Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; GT 8 Informação e Tecnologia; GT 9 Museu, Patrimônio e Informação; GT 10 Informação e Memória e GT 11 Informação & Saúde.

³ PHP é um acrônimo recursivo para PHP: *Hypertext Preprocessor*.

simplicidade do programa, sua criação e desenvolvimento contribuíram sobremaneira para a efetivação dos objetivos da investigação.

Algumas intervenções, como correções e/ou ajustes, foram executadas nas palavras-chave coletadas, a fim de aplicar uma padronização que trouxesse resultados mais confiáveis pelo SEP. Essa etapa, denominada normalização dos dados, ocorreu sob comandos executados no SEP, do seguinte modo:

- Adequação de plurais: palavras como, por exemplo, informação/informações ou documento/documentos, foram alteradas para a forma singular, o que possibilitou seu agrupamento para fins quantitativos, tornando-as igualitárias;
- Padronização de preposições: termos como "recuperação da informação" e "recuperação de informação", "política de informação" e "política da informação", por exemplo, foram alteradas para uma única variação. Adotou-se a forma constante no TBCI, quando possível, e, nos demais casos, optou-se pela forma mais recorrente; e
- Padronização de letras maiúsculas e minúsculas: termos como *Web Semântica*, *WEB Semântica* ou *Web semântica* foram agrupados em uma única forma, com maior número de ocorrências.

Para a etapa de verificação do controle de vocabulário de palavras-chave, o estudo se fundamentou no vocabulário controlado Tesouro Brasileiro de Ciência da Informação (TBCI), com o propósito de comparar a aplicabilidade do instrumento nas palavras-chave atribuídas pelos pesquisadores do ENANCIB, em suas publicações.

O TBCI foi finalizado, em 2013, por Pinheiro e Ferrez (2014), além de outros pesquisadores e técnicos, os quais, juntos, trabalharam para a concretização do tesouro. Esse tesouro possui versão gratuita e em PDF, com mais de 1.800 termos (a maioria com versões em inglês e espanhol), complementados por definições (SUAIDEN, 2014). Atualmente, o TBCI pode ser consultado na *Web*⁴, em um formato mais dinâmico, que lembra um mecanismo de busca e possui suporte para a navegação por categorias e assuntos. Essa

⁴ Disponível em: <http://www.uel.br/revistas/informacao/tbci/vocab/sobre.php>. Acesso em: 12 ago. 2020.

interface de acesso é fruto de um projeto desenvolvido em parceria entre a Universidade Estadual de Londrina (UEL), Universidade Estadual “Júlio de Mesquita Filho” (UNESP), Câmpus de Marília, e Instituto Brasileiro de Ciência e Tecnologia (IBICT). A parceria entre essas instituições tornou possível o acesso ao TBCI, em um *site* que permite buscar termos em barra de navegação (PINHEIRO; FERREZ, 2014).

A coleta dos dados foi iniciada através do acesso aos *sites* das edições do ENANCIB, de 2015 a 2018, em busca dos seus respectivos Anais. Na sequência, os arquivos em formato PDF dos trabalhos foram categorizados, de acordo com o ano e GT em que foram publicados. No total, 1.075 artigos foram analisados, na pesquisa, distribuídos conforme se observa na Tabela 1.

Tabela 1 – Quantidade de trabalhos analisados por ano e GT

GT	2015	2016	2017	2018	Total
GT 1 - Estudos Históricos e Epistemológicos da Ciência da Informação	14	23	22	21	80
GT 2 - Organização e Representação do Conhecimento	25	49	44	36	154
GT 3 - Mediação, Circulação e Apropriação da Informação	19	22	27	37	105
GT 4 - Gestão da Informação e do Conhecimento	30	28	39	36	133
GT 5 - Política e Economia da Informação	13	19	25	35	92
GT 6 - Informação, Educação e Trabalho	11	19	18	17	65
GT 7 - Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação	20	32	30	28	110
GT 8 - Informação e Tecnologia	26	23	33	31	113
GT 9 - Museu, Patrimônio e Informação	21	16	18	28	83
GT 10 - Informação e Memória	24	32	18	28	102
GT 11 - Informação & Saúde	14	9	8	7	38
Total por ano	217	272	282	304	
Total de artigos da amostra					1.075

Fonte: Elaborado pelos autores.

A partir da Tabela 1, fica perceptível que, a cada nova edição do ENANCIB, mais trabalhos são publicados, com recorde no ano de 2018, com 304 artigos. Outro ponto notório nas estatísticas apresentadas nesta compilação dos dados é que o GT 2 foi aquele que mais publicou em relação aos demais grupos de trabalho. Assim, o GT 2 teve 154 publicações, nos anos da amostra, contra apenas 38 trabalhos do GT 11, o qual foi o que menos publicou. A saber, talvez a tradição e a robustez do GT 2, existente desde 1995, tenha contribuído para esse resultado. O GT 11, inaugurado em 2011, é um grupo recente do

ENANCIB e ainda caminha para a sua consolidação.

Em decorrência da pesquisa, o próximo passo foi extrair as palavras-chave dos 1.075 artigos que compuseram a amostra com o SEP, *software* desenvolvido para que a mineração das palavras-chave ocorresse de maneira automática.

A normalização dos dados possibilitou os ajustes necessários, para que os resultados coletados ficassem uniformes e padronizados. Na etapa posterior, relativa à tabulação dos dados, as palavras-chave foram postas em ordem decrescente do número de ocorrências para cada GT e ano. Apenas as três palavras com maior número de ocorrências foram consideradas, cujo recorte se baseou na recomendação da quantidade de palavras-chave permitidas, durante as submissões dos trabalhos completos no ENANCIB, que é de três a cinco palavras, no máximo.

O próximo passo foi verificar e comparar as palavras-chave das publicações dos Anais do ENANCIB, nos anos da amostra (2015 a 2018), com os termos do TBCI. A comparação entre as palavras-chave com o TBCI foi realizada de modo manual, com o tesauro em PDF disponibilizado pelo IBICT e em sua versão *on-line*, no *site*. Desse modo, as palavras-chave recuperadas nas publicações do ENANCIB foram comparadas uma a uma, no TBCI. O interesse, nessa etapa, foi investigar o uso do TBCI para a representação documentária, pelos pesquisadores que definem suas palavras-chave, no Portal do evento.

Os resultados obtidos na análise e comparação das palavras-chave utilizadas pelos pesquisadores com o TBCI foram classificados como “muito satisfatório”, “satisfatório”, “pouco satisfatório”, “insatisfatório”, “muito insatisfatório” e “totalmente insatisfatório”, conforme segue:

- **GT 4 = 100% - muito satisfatório;**
- GT 2, 6, e 8 = 83,3 % - satisfatório;
- GT 5 e 7 = 66,6% - pouco satisfatório;
- GT 1, 9 e 11 = 50% - insatisfatório;
- GT 3 = 25,0% - muito insatisfatório;
- **GT 10 = 8,3% - totalmente insatisfatório.**

4 RESULTADOS

Os resultados da análise sobre o controle de vocabulário das palavras-chaves atribuídas pelos autores de trabalhos completos, em metadados de assuntos, durante o autoarquivamento no sistema de evento científico, são apresentados, a seguir, nos dois itens de diretrizes aos autores e extração e mineração das palavras-chave, por meio da criação do *Software* de Extração de Palavras-Chave (SEP).

4.1 ANÁLISE DAS DIRETRIZES AOS AUTORES SOBRE CONTROLE DE VOCABULÁRIO DE PALAVRAS-CHAVE

A escolha das palavras-chave é feita pelo próprio pesquisador, ao submeter seu trabalho ao ENANCIB. A partir da análise das orientações sobre o controle de vocabulário de palavras-chave, não foi identificada nenhuma influência ou validação de um profissional da informação externo, nesse processo, assim como não foram encontrados manual ou guia de indexação ou mesmo uma política de indexação capaz de auxiliar os pesquisadores, na escolha de suas palavras-chave.

A validação das palavras-chave atribuídas pelos autores e inseridas na submissão dos trabalhos completos ao ENANCIB se dá no processo de avaliação por pares, momento em que a qualidade e a representatividade das palavras-chave são consideradas, em função de um rigoroso processo de avaliação.

Com foco principalmente nas Diretrizes para os autores, item disponível no *site* do ENANCIB direcionado às regras para publicação no evento, foi possível identificar as ações que refletem a intenção do ENANCIB em constituir um controle do vocabulário, em suas publicações. A primeira ação observada foi a recomendação pelo uso do Tesouro Brasileiro de Ciência da Informação (TBCI), quando na definição das palavras-chave utilizadas pelos pesquisadores, de sorte a representar suas pesquisas e trabalhos submetidos ao evento. Por refletir a configuração científica da Ciência da Informação, na contemporaneidade, conforme mandala representativa proposta por Pinheiro

(2018), o uso do TBCI na definição de palavras-chave propicia aos trabalhos expostos no ENANCIB consistência de terminologia e de vocabulário, aspectos fundamentais para a recuperação da informação. De forma clara, a indicação do tesauro na definição de palavras-chave, no Portal do evento, ocorre da seguinte maneira:

Palavras-chave/Keywords: incluir entre 3 (três) e 5 (cinco) palavras-chave/keywords, preferencialmente de acordo com o Tesauro Brasileiro de Ciência da Informação das autoras Lena Vânia Ribeiro Pinheiro e Helena Dodd Ferrez [...] resguardando o uso de termos específicos de cada área, quando estes não forem contemplados pelo referido tesauro. (ENANCIB, 2015-2018, não paginado).

Com base na recomendação constante no Portal do ENANCIB, foi constatado que há indicação de utilização do TBCI, mas não há obrigatoriedade de seu uso. Cintra *et al.* (2002, p. 41) explicam que, “[...] conseqüentemente, abandona-se a pretensão de cobrir todo o conhecimento para voltar-se a domínios cada vez mais específicos.” Apesar disso, as contribuições do TBCI, para auxiliar no controle de vocabulário da Ciência da Informação e suas áreas, são indiscutíveis.

Outro ponto analisado foi a sugestão do emprego dos termos específicos de cada área dos 11 GTs, quando algum termo não for encontrado no TBCI. Entretanto, nada mais a respeito é esclarecido, nesse sentido. De toda maneira, a indicação de uso do TBCI para a definição das palavras-chave pelos pesquisadores foi a única ação encontrada voltada à representação documentária, nos Anais do evento, de maneira mais concreta. A recomendação para o uso de vocabulário controlado é um apoio importante aos pesquisadores, para a representação do conteúdo de suas pesquisas por palavras-chave, as quais, quando bem aplicadas, conferem padronização, precisão e qualidade para a pesquisa e a recuperação temática.

Todavia, não fica claro se as atualizações sofridas pelo TBCI *on-line* são referentes aos termos do tesauro ou às atualizações necessárias a recursos digitais. Tampouco ficou evidente se a versão no *site* possui realmente 2.058 termos (como sugerido, na plataforma), em contraponto aos 1.800 da versão em PDF. De todo modo, o TBCI foi o vocabulário controlado escolhido para a aplicação da pesquisa, já que fora anteriormente apontado aos pesquisadores

do ENANCIB como forma de proverem controle de vocabulário de suas pesquisas.

4.2 ANÁLISE DO CONTROLE DE VOCABULÁRIO DE PALAVRAS-CHAVE POR MEIO DE EXTRAÇÃO E MINERAÇÃO

Os resultados da análise são apresentados nas Tabelas 2, 3 e 4. Em cada um deles, é possível visualizar as três palavras-chave mais recorrentes em cada ano e GT demarcadas por “Consta no TBCI”, “Não consta no TBCI” e “Termo não preferido”.

Tabela 2 – As 3 palavras-chave mais recorrentes no GT 1, GT 2, GT 3 e GT 4

GT 1 - Estudos Históricos e Epistemológicos da Ciência da Informação								
Ranking	Qtde	2015	Qtde	2016	Qtde	2017	Qtde	2018
1ª	4	Informação	6	Ciência da Informação	7	Ciência da Informação	10	Ciência da Informação
2ª	4	Ciência da Informação	4	Informação	3	Documentação	3	Informação
3ª	2	Ética	3	Interdisciplinaridade	2	Documento	3	Interdisciplinaridade
Palavra-chave mais recuperada na amostra: Ciência da Informação (27 ocorrências)								
GT 2 - Organização e Representação do Conhecimento								
Ranking	Qtde	2015	Qtde	2016	Qtde	2017	Qtde	2018
1ª	4	Organização e representação do conhecimento	12	Organização do Conhecimento	9	Organização do Conhecimento	6	Organização do Conhecimento
2ª	2	Representação	6	Ontologia	4	Ontologia	3	Sistemas de Organização do Conhecimento
3ª	2	Ontologia	6	Semiótica	4	Representação do Conhecimento	3	Indexação
Palavra-chave mais recuperada na amostra: Organização do Conhecimento (27 ocorrências)								
GT 3 - Mediação, Circulação e Apropriação da Informação								
Ranking	Qtde	2015	Qtde	2016	Qtde	2017	Qtde	2018
1ª	5	Mediação da Informação	5	Mediação da informação	6	Mediação da Informação	6	Competência em informação
2ª	2	Competência	3	Mediação	3	Mediação	5	Mediação cultural

		informacion al						
3ª	2	Competênci a em Informação	3	Arquivologia	3	Competênci a Informacion al	3	Mediação da informação
Palavra-chave mais recuperada na amostra: Mediação da Informação (19 ocorrências)								
GT 4 - Gestão da Informação e do Conhecimento								
Ranking	Qt de	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1ª	4	Gestão do conhecimen to	7	Gestão da informação	12	Gestão do Conhecime nto	8	Gestão do Conhecimento
2ª	4	Inteligência competitiva	7	Gestão do conhecimento	8	Gestão da informação	7	Gestão da informação
3ª	3	Gestão da informação	3	Bibliotecas Universitárias	4	Competênci a em Informação	3	Inovação
Palavra-chave mais recuperada na amostra: Gestão da informação (25 ocorrências)								
Legenda:		Consta no TBCI		Não consta no TBCI				Termo não preferido

Fonte: Elaborado pelos autores.

Como é explícito no Tabela 2, no GT 1 – Estudos Históricos e Epistemológicos da Ciência da Informação, a palavra-chave com maior recorrência, em todo o período, foi **Ciência da Informação**, seguida de **Informação**. Apesar de essas palavras serem de total importância para a Ciência da Informação, uma vez que a primeira demonstra o próprio nome do campo e a segunda, seu principal objeto de estudo (CAPURRO; HJØRLAND, 2007; PINHEIRO, 2004), ambas são palavras com pouca significação, quando consideradas individualmente, na recuperação da informação. Isso significa que, na recuperação da informação, os termos retornariam muitos resultados e teriam alta revocação, porém, com pouca ou nenhuma precisão.

Quanto ao TBCI, somente 50% das 12 palavras-chave listadas estavam em concomitância com o tesouro, o que representa, para o presente estudo, um resultado tido como insatisfatório. Ainda assim, a maioria das palavras-chave isoladas pouco significam, em uma estratégia de busca. É o caso das palavras "Interdisciplinaridade" ou "Documento", por exemplo. Já a palavra-chave "Documentação" existe no TBCI como um termo não preferido, demonstrando que não seria a palavra ideal para representar adequadamente a informação. Entretanto, seria uma opção viável, como termo controlado, a fim de representar a temática, conforme o tesouro.

O GT 2 - Organização e Representação do Conhecimento – revela 83,3% de concomitância de suas palavras-chave, quando comparadas com o tesauro, o que é considerado um resultado satisfatório e capaz de indicar maturidade dos pesquisadores desse grupo, na atribuição das palavras-chave que representam seus trabalhos. Uma das palavras que não pertencem ao tesauro e que teve quatro ocorrências, no ano de 2015, foi "Organização e Representação do Conhecimento", a qual, além de constituir o próprio nome do GT, representa uma das ramificações basilares da Ciência da Informação, no contexto da própria Organização do Conhecimento. Por esses motivos, e dada a importância da área para a Ciência da Informação, conforme já explicitado por Brasher e Café (2008) e Vignoli, Souto e Cervantes (2013), causa alguma estranheza o termo composto não constar no TBCI.

De toda maneira, nos anos posteriores, a palavra-chave mais recorrente do GT 2 foi "Organização do Conhecimento", que é um termo preferido no TBCI e expressa a fidedignidade dos pesquisadores ao GT e à sua grande área de pesquisa, no escopo da Ciência da Informação.

Uma situação distinta ocorre no GT 3 – Mediação, Circulação e Apropriação da Informação, no qual, das 12 palavras-chave mais recorrentes, somente cinco figuram no TBCI, sendo que duas delas são termos não preferidos. Dessa forma, somente 25% das palavras-chave mais recorrentes desse grupo pertencem ao TBCI. Assim, pode-se depreender que o GT teve um desempenho e resultados muito insatisfatórios, na representação documentária de seus termos.

O GT 4 – Gestão da Informação e do Conhecimento – teve 100% de concomitância de suas palavras-chave com o TBCI, um resultado classificado como muito satisfatório, visto que todas as palavras-chave recuperadas corresponderam ao seu termo preferido, no TBCI. Tal fato pode apontar que há, pelos pesquisadores desse grupo, um excelente entendimento sobre a função das palavras-chave, na representação de suas pesquisas, além de conhecimento da existência do TBCI.

Tabela 3 – As 3 palavras-chave mais recorrentes no GT 5, GT 6, GT 7 e GT 8

GT 5 - Política e Economia da Informação								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1 ^a	5	Lei de Acesso à Informação	3	Lei de Acesso à informação	3	Lei de Acesso à Informação	6	Regimes de informação
2 ^a	2	Biblioteca escolar	2	Biblioteca Pública	3	Gestão de documentos	4	Lei de Acesso à Informação
3 ^a	2	Gestão de documentos	2	Regimes de informação	2	Regimes de informação	4	política de informação
Palavra-chave mais recuperada na amostra: Lei de Acesso à Informação (15 ocorrências)								
GT 6 - Informação, Educação e Trabalho								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1 ^a	2	Tecnologias de Informação e Comunicação	5	Competência em informação	2	Informação	4	Ciência da Informação
2 ^a	2	Regimes de informação	3	Bibliotecas Universitárias	2	Busca de Informação	3	Competência em informação
3 ^a	2	Biblioteconomia	2	Atuação Profissional	2	Uso da informação	2	Arquivologia
Palavra-chave mais recuperada na amostra: Competência em informação (8 ocorrências)								
GT 7 - Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1 ^a	5	Comunicação científica	8	Comunicação científica	4	Comunicação científica	7	Ciência da Informação
2 ^a	5	Produção científica	5	Produção científica	3	Produção científica	6	Comunicação científica
3 ^a	4	Cientometria	4	Ciência da Informação	2	Periódico científico	6	Produção científica
Palavra-chave mais recuperada na amostra: Comunicação científica (23 ocorrências)								
GT 8 - Informação e Tecnologia								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1 ^a	3	Web Semântica	4	Recuperação da informação	7	Web Semântica	3	Informação e Tecnologia
2 ^a	3	Acessibilidade	3	Informação e Tecnologia	4	Metadados	3	Web Semântica
3 ^a	2	Recuperação da informação	3	Ontologia	3	Ontologia	3	Arquitetura da Informação
Palavra-chave mais recuperada na amostra: Web Semântica (13 ocorrências)								
Legend a:		Consta no TBCI		Não consta no TBCI		Termo não preferido		

Fonte: Elaborado pelos autores.

De acordo com o Tabela 3 e nos resultados do GT 5 – Política e Economia da Informação, foi notado um fato interessante: a palavra-chave mais recorrente é "Lei de Acesso à Informação". A LAI (Lei de Acesso à Informação) desponta entre as três primeiras palavras-chave, em todos os anos, além de aparecer por 15 vezes no total nos anos pesquisados. Esse resultado totaliza 66,6% de concomitância do GT com o tesouro, o que o classifica como pouco satisfatório, nos preâmbulos deste estudo.

No entanto, o termo "Lei de Acesso à Informação" não foi encontrado no TBCI, exprimindo que talvez o tesouro careça de atualizações, uma vez que a LAI ou Lei nº 12.527, criada no ano de 2011, para regulamentar o acesso à informação, previsto no inciso XXXIII do artigo 5º da Constituição (BRASIL, 2011), é assunto preponderante em investigações desenvolvidas pelos pesquisadores vinculados ao GT

No GT 6 – Informação, Educação e Trabalho, tem-se uma percentagem de concomitância de 83,3%, revelando que a definição das palavras-chave pelos pesquisadores foi bastante adequada e que os resultados desse GT foram considerados satisfatórios. Uma das palavras-chave que não pertence ao TBCI é "Informação", a qual, como já foi discutido, apesar de ser o principal objeto de estudo da CI, não é um termo adequado para fins de recuperação da produção científica.

Observa-se, no GT 7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação, uma percentagem de concomitância de termos com o TBCI de 66,6%, um resultado entendido como pouco satisfatório. A palavra-chave "Produção Científica" emerge, nos quatro anos, entre as palavras-chave mais empregadas, mas não pertence ao TBCI. Entretanto, há, no TBCI, o termo "Produtividade Científica", o qual aparece como termo relacionado à "Comunicação Científica", que, não por acaso, é o termo mais recorrente no próprio GT. De acordo com Pinheiro e Ferrez (2014, p. 49), o termo "Comunicação Científica" é definido no tesouro por "[...] todo espectro de atividades associadas com a produção, disseminação e uso da informação." Isso parece fazer transparecer que há uma possível repetição de termos desnecessários, já que o termo preferido é "Comunicação Científica."

No GT 8 – Informação e Tecnologia, foram observadas duas ocorrências de seu título, como palavra-chave, em anos diferentes, sendo que "Informação e Tecnologia" é um termo não pertencente ao TBCI. Porém, acredita-se que o termo poderia ser substituído por seu termo preferido, "Tecnologias da Informação e Comunicação" disposto no tesauro. De acordo com a definição do TBCI, "Tecnologias da Informação e Comunicação" é um termo que "[...] abarca todas as formas de tecnologias usadas para criar, armazenar, intercambiar e usar a informação em suas várias formas." (PINHEIRO; FERREZ, 2014, p. 197). Assim, a percentagem de concomitância das palavras-chave recuperadas nesse GT, quando comparadas aos termos do TBCI, foi de 83,3%, o que é considerado um resultado satisfatório.

Tabela 4 – As 3 palavras-chave mais recorrentes no GT 9, GT 10 e GT11

GT 9 – Museu, Patrimônio e Informação								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1ª	8	Museologia	2	Musealização	7	Museus	10	Museologia
2ª	7	Museus	2	Patrimonialização	6	Museologia	6	Museus
3ª	4	Patrimônio	2	Patrimônio Cultural Imaterial	4	Patrimônio	3	Patrimônio Cultural
Palavra-chave mais recuperada na amostra: Museologia (24 ocorrências)								
GT 10 - Informação e Memória								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1ª	12	Memória	18	Memória	9	Memória	11	Memória
2ª	6	Informação	8	Informação	2	Patrimônio	3	Patrimônio Cultural
3ª	2	Patrimônio cultural	4	Memória social	2	Arquivo Pessoal	3	Ciência da Informação
Palavra-chave mais recuperada na amostra: Memória (50 ocorrências)								
GT 11 - Informação & Saúde								
Ranking	Qtd e	2015	Qtd e	2016	Qtd e	2017	Qtd e	2018
1ª	3	Bibliometria	1	Ética da Informação	1	Análise de domínio	2	Acesso à informação
2ª	1	Citações bibliográficas	1	Fonte de Informação Primária	1	Comportamento informacional	2	Informação em saúde

3ª	1	Estudos de citação	1	Prontuário de Paciente	1	Equipe multiprofissional hospitalar	2	Ciências da Saúde
Palavra-chave mais recuperada na amostra: Bibliometria (3 ocorrências)								
Legend a:		Consta no TBCI		Não consta no TBCI		Termo não preferido		

Fonte: Elaborado pelos autores.

Segundo o Tabela 4, no GT 9 – Museu, Patrimônio e Informação, foi possível observar uma concomitância de 50% das palavras-chave com o TBCI, resultando em uma classificação insatisfatória. Dessas palavras-chave, cinco lidam com o conceito de "Patrimônio", termo que não foi encontrado no tesauro. Nenhum outro termo que pudesse substituir essa palavra foi localizado no TBCI. A mesma situação se dá com a palavra-chave "Musealização", a qual, para Schärer (2009, p. 87), significa “[...] preservação dos valores ideais das coisas como sinais [...]”, enquanto, segundo Davallon (2009, p. 15, tradução nossa) é entendida como “[...] a transformação da situação de um objeto em um objeto museológico.” Com base nas definições, não foi possível localizar, no TBCI, um termo adequado que pudesse substituir a atribuição anteriormente efetuada pelos pesquisadores. Tais fatos reiteram a necessidade de adequação do tesauro para certas áreas, como a Museologia, que, apesar de contradições em pertencer ou não à Ciência da Informação (ARAÚJO, 2014), ainda segue próxima à Ciência da Informação, nas avaliações da Fundação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - (2017).

Quantitativamente, o GT 10 – Informação e Memória – apresenta o pior resultado, dentre os GTs da amostra, com 8,3% de concomitância de suas palavras-chave com o tesauro. Por esse motivo, o resultado foi categorizado como totalmente insatisfatório. Nesse GT, a palavra-chave "Ciência da Informação" foi o único termo encontrado no TBCI. Porém, conforme já discutido anteriormente, ela não é um bom descritor para representar os documentos, tendo em vista uma estratégia de busca. Quanto às demais palavras-chave, "Informação", "Patrimônio" e "Memória", apesar de possuírem o maior número de ocorrências, dentre todos os GTs da amostra, não são termos pertencentes ao TBCI.

Segundo Boccato e Fujita (2010), linguagens documentárias são

construídas com base na terminologia de uma área, de sorte a possibilitar representação e recuperação da informação adequada. Entretanto, sem a padronização dos termos, a recuperação da informação, nos textos do GT 10, provavelmente seria difícil. Por isso, evidenciam Silva e Fujita (2004, p. 138): “O bom ou mau desempenho da indexação reflete-se na recuperação da informação feita pelos índices.”

Quanto ao GT 11 – Informação & Saúde, o qual obteve 50% de concomitância com o TBCI, apenas cinco palavras-chave constam no tesauro, embora os termos não pareçam representar, de fato, o núcleo de pesquisa do GT, o que foi demarcado como um resultado insatisfatório. Um exemplo disso é a palavra-chave com maior número de ocorrências (três vezes), durante a análise, que é "Bibliometria." Sabe-se que a Bibliometria pode ser aplicada em qualquer tipo de estudo da Ciência da Informação, assim como em qualquer GT do ENANCIB. No entanto, não é um termo substancial em relação aos termos preponderantes nos estudos de Informação e Saúde. Assim, compreende-se que, apesar de algumas palavras-chave atribuídas pelos pesquisadores pertencerem ao TBCI, esse GT teve igualmente um dos resultados mais insatisfatórios, associado à prática da representação documentária através de controle de vocabulário.

Ressalta-se, ainda, que a palavra-chave "Informação em saúde" não possui uma alternativa substitutiva adequada, no tesauro, o que sinaliza novamente a uma possível necessidade de atualização do vocabulário controlado, para contemplar os assuntos relativos à área da Saúde. A palavra-chave "Prontuário do paciente", termo importante no contexto dos estudos de informação e saúde, possui uma palavra-chave com significado próximo, que é "Prontuário médico." Um prontuário médico, de acordo com a Resolução nº 1.638/2002 do Conselho Federal de Medicina (2002, p. 184-185), é “[...] o documento único constituído de um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada.” Sob análise da definição do CFM, cabe ressaltar que possivelmente poderia acontecer a substituição pelo termo existente no tesauro ("prontuário médico"), sem grandes problemas linguísticos,

mas isso não aconteceu. Nesse caso, observa-se que a comunidade científica prefere o termo não constante no TBCI, para indexar suas pesquisas na base de dados do evento.

De toda forma, ao ponto que alguns GTs se saíram muito bem na representação documentária, conforme o TBCI, como o GT 4 e os GT 2, 6 e 8, o primeiro com 100% de concomitância e os demais com 83,3%, outros foram um tanto quanto deficientes, na comparação com o tesouro, como o GT 11, o qual obteve acertos de somente 50%, e o GT 10, com apenas 8,3%. Esse resultado denota a urgência de os pesquisadores integrantes dos GTs com os piores resultados da análise padronizarem seus termos por linguagens de indexação, como as constantes no TBCI.

De acordo com as percentagens obtidas em cada GT, a média geral de utilização do tesouro foi de 60,5%. Essa percentagem é relevante, pois mostra que os pesquisadores utilizaram o TBCI e realizaram o controle de vocabulário. Entretanto, ainda são necessárias mais ações voltadas à prática da representação documentária, com o uso do tesouro, para que o controle do vocabulário nos Anais do ENANCIB assegure com mais precisão e correção a recuperação da informação.

Os resultados revelam ainda que muitas das palavras-chave empregadas na indexação dos trabalhos ainda são deficitárias, quanto à padronização pelo TBCI. Não se sabe ao certo se elas são relacionadas a termos específicos da área de pesquisa dos GTs correspondentes ou se são apenas palavras vazias, as quais nada significam, em uma estratégia de busca. É, portanto, primaz avaliar o uso de vocabulários controlados, a fim de representar as pesquisas do ENANCIB.

Se um tesouro constitui uma lista de termos controlados, os quais devem representar conceitos de um domínio do conhecimento, além das estruturas hierárquicas com relações semânticas (MOREIRO GONZÁLEZ, 2011), era esperado que os termos do TBCI fossem mais adotados como palavras-chave, nos trabalhos completos apresentados no ENANCIB. Em atenção às transformações da Ciência da Informação, desde a publicação do TBCI, as diretrizes do XXIII ENANCIB resguardam o uso de termos específicos pelos

pesquisadores, para a atribuição de palavras-chave, quando estes não forem contemplados pelo TBCI. Importante enfatizar que o uso de vocabulários controlados do tipo tesouro, para a atribuição de palavras-chave pelo pesquisador, confere rigor terminológico à representação do conteúdo e conceitos científicos contemplados na pesquisa.

Sob esses aspectos, acredita-se que tais resultados representem a necessidade de melhorias na representação documentária e controle de vocabulário, nas publicações dos eventos, além de incentivar sua prática por pesquisadores da Ciência da Informação, essencialmente nas próximas edições do ENANCIB. Igualmente, os resultados evidenciam a urgência de atualizações nos termos do TBCI, como forma de abarcar termos e tendências mais recentes, presentes nas pesquisas atuais da Ciência da Informação.

5 CONCLUSÕES

O controle do vocabulário, na atribuição de palavras-chave de trabalhos completos, publicados nos Anais do ENANCIB dos anos de 2015 a 2018, foi analisado e comparado ao Tesouro Brasileiro de Ciência da Informação (TBCI), após extração e mineração com o *Software* de Extração de Palavras-Chave (SEP), desenvolvido para essa finalidade. A partir dessa análise, foi identificado que a indexação de trabalhos completos do ENANCIB ocorre por meio da atribuição de palavras-chave promovida pelos próprios pesquisadores, e que cabe aos pareceristas avaliar a qualidade e a representatividade das palavras-chave atribuídas.

Foi verificado que 60,5% das palavras-chave têm idêntica correspondência com os termos do TBCI, enquanto 39,5% das palavras-chave utilizadas pelos pesquisadores, na indexação de seus trabalhos, não estão em consonância com o TBCI indicado aos pesquisadores para adoção, na atribuição das palavras-chave dos seus textos.

Dessa forma, um questionamento é levantado: qual a origem dessas palavras usadas como palavras-chave pelos pesquisadores do ENANCIB? Como não foi possível identificar respostas para a questão, acredita-se que há outro nicho para novas pesquisas acerca da representação documentária no

ENANCIB, com base neste estudo.

Outro ponto de destaque, na análise dos resultados, foi a sobressaliência de acertos no uso das palavras-chave de um GT a outro, a qual chegou ao máximo absoluto de 100%, no GT 4, contra um mínimo preocupante de 8,3%, no GT 10. Embora o GT 11 tenha obtido resultados mais satisfatórios, na concomitância das palavras-chave, em relação ao GT 10, seus termos recuperados e analisados pouco representam a área da Saúde e, por isso, a representação documentária e o controle de vocabulário, nesse GT, também foram bastante deficitários.

Os dados coletados demonstram que mudanças são necessárias, para que os trabalhos que compõem os Anais do ENANCIB possuam um controle de vocabulário adequado e factível por seus pesquisadores. Acredita-se, de igual maneira, que as ações voltadas ao estabelecimento de um controle de vocabulário, nos Anais do ENANCIB, sejam ainda incipientes e pouco divulgadas. Nesse sentido, ressalta-se que a indicação de utilização do TBCI para a definição de palavras-chave pelos pesquisadores do evento demorou a acontecer (2015), uma vez que o TBCI foi publicado em 2008, sendo, inclusive, divulgado em pesquisas no ENANCIB. Recomenda-se a elaboração de tutoriais aos autores, com orientações e benefícios sobre o uso do TBCI para controle de vocabulário.

Como pesquisa futura relacionada a esta, pretende-se identificar como ocorre a recuperação da informação, no Portal do ENANCIB, sob as palavras-chave já indexadas nos Anais, pois, conforme é sabido, se a representação documentária não é bem realizada, por meio de controle de vocabulário e uso de um vocabulário controlado, a recuperação da informação provavelmente será prejudicada.

REFERÊNCIAS

ALMEIDA, T. de. **Os loci epistêmicos e o método analítico como forma de compreensão do ensino e da pesquisa em organização do conhecimento no Brasil do século XXI**. 2019. 422 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, Instituto Brasileiro de

Informação em Ciência e Tecnologia, Rio de Janeiro, 2019. Disponível em: <http://ridi.ibict.br/handle/123456789/1046>. Acesso em: 06 de julho de 2023.

ARAÚJO, C. A. A. **Arquivologia, Biblioteconomia, Museologia e Ciência da Informação**: o diálogo possível. Brasília, DF: Briquet de Lemos/Livros, São Paulo, SP: Associação Brasileira de Profissionais da Informação (ABRAINFO), 2014. 215p.

BOCCATO, V. R. C.; FUJITA, M. S. L. O uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias: avaliação sociocognitiva com protocolo verbal. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 15, n. 3, 23-51, 2010.

BRASCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento? *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 9., 2008, São Paulo. **Anais [...]**. São Paulo, SP: Escola de Comunicação e Artes (ECA) USP; ANCIB, 2008. Disponível em: <http://www.ancib.org.br/media/dissertacao/1835.pdf>. Acesso em: 02 nov. 2020.

BRASIL. **Lei n.º 12.527, de 18 de novembro de 2011**. Dispõe sobre os procedimentos a serem observados pela União, Estados, Distrito Federal e Municípios, com o fim de garantir o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal. Diário Oficial [da] República Federativa do Brasil, seção 1. Brasília: Presidência da República, 2011.

CAPURRO, R.; HJØRLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 148-207, jan./jun. 2007.

CINTRA, A. M.; TÁLAMO, M. de F. G. M.; LARA, M. L. G. de; KOBASHI, N. Y. **Para entender as linguagens documentárias**. 2. ed. rev. e ampl. São Paulo: Polis, 2002. (Coleção Palavra-Chave, 4).

CONSELHO FEDERAL DE MEDICINA (CFM). **Resolução nº 1.638, de 10 de julho de 2002**. Define prontuário médico e torna obrigatória a criação da Comissão de Revisão de Prontuários nas instituições de saúde. Brasília, Seção I, p. 184-5, 2002.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **Tabelas de áreas do conhecimento**. Brasília: CAPES, 2017.

COSTA, M. U. P.; MOURA, M. A. A representação da informação em contextos de comunicação científica: a elaboração de resumos e palavras-chave pelo pesquisador-autor. **Informação & Informação**, Londrina, v. 18, n. 3, p. 45-67, 2013.

DAVALLON, J. Heritage, preservation, research, object, collection, musealization: session 4. *In*: ANNUAL COFOM SYMPOSIUM. MUSEOLOGY:

BACK TO BASICS; SYNTHESIS OF THE SYMPOSIUM, 1., Mariemont, 2009. **Proceedings** [...]. Liège: Mariemont, 2009. p. 12-23.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 15., 2015, João Pessoa. **Anais** [...]. João Pessoa: UFPB, ANCIB, 2015. Disponível em:
<http://www.ufpb.br/evento/index.php/ENANCIB2015/ENANCIB2015/schedConf/presentations>. Acesso em: 25 out. 2020.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 17., 2016, Salvador. **Anais** [...]. Salvador: ANCIB, UFBA, ANCIB, 2016. Disponível em:
https://drive.google.com/file/d/0B7rxeg_cwHajMW9ZV0xFZHBhTnc/view. Acesso em: 12 out. 2020.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 18., 2017, Marília. **Anais** [...]. Marília: UNESP, ANCIB, 2017. Disponível em:
<http://ENANCIB.marilia.unesp.br/index.php/xviiiENANCIB/ENANCIB/schedConf/presentations>. Acesso em: 12 set. 2020.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 19., 2018, Londrina. **Anais** [...]. Londrina: UEL, ANCIB, 2018. Disponível em:
<http://ENANCIB.marilia.unesp.br/index.php/XIXENANCIB/xixENANCIB/schedConf/presentations>. Acesso em: 12 jul. 2020.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 20., 2019, Florianópolis. **Anais** [...]. Florianópolis: UFSC, ANCIB, 2019. Disponível em:
<http://ENANCIB.marilia.unesp.br/index.php/XIXENANCIB/xixENANCIB/schedConf/presentations>. Acesso em: 12 set. 2020.

FERREIRA, A. C.; MACULAN, B. C. M. dos S. Fundamentos teórico-conceituais aplicáveis à revisão e à atualização de tesouros. **Ciência da Informação em Revista**, Maceió, v. 7, n. 1, p. 46-70, jan./abr. 2020.

FUJITA, M. S. L.; AGUSTÍN-LACRUZ, M. Del C.; TERRA, A. L. S. Journals' guidelines about title, abstract and keywords: an overview of Information Science and Communication Science areas. **Journal: European Science Editing**, v. 44, p. 76-79, 2018.

FUJITA, M. S. L.; TARTAROTTI, R. C. D. Análise de palavras-chave da produção científica de pesquisadores: o autor como indexador. **Informação & Informação**, v. 25, n. 3, p. 332-374, 2020.

GROSS, T.; TAYLOR, A. G.; JOUDREY, D. N. Still a lot to lose: the role of controlled vocabulary in keyword searching. **Cataloging & Classification Quarterly**, [S. l.], v. 53, n. 1, p. 1-39, 2015.

HARTLEY, J.; KOSTOFF, R. N. How useful are 'keywords' in scientific journals? **Journal of Information Science**, [S. l.], v. 29, n. 5, p. 433-438, 2003.

HJØRLAND, B. Does the traditional thesaurus have a place in modern information retrieval? **Knowledge Organization**, Germany, v. 43, n. 3, p. 145-159, 2016.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964-1: 2011** Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization, 2011.

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão de literatura. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

MARTINS, G. K.; MOTA, D. A. R.; FELIPE, A. A. C. Tendências em pesquisa no GT2 do ENANCIB: análise das comunicações científicas de 2003 a 2016 com base nas palavras-chave atribuídas pelos autores. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 18., 2017, Marília. **Anais [...]**. Marília: Unesp, 2017. p. 1-22. Disponível em: http://enancib.marilia.unesp.br/index.php/XVIII_ENANCIB/ENANCIB/paper/view/579. Acesso em: 06 de julho de 2023.

MORAES, R. P. T. de; CAMPOS, M. L. de Mapeamento temático: o GT2 do ENANCIB. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais [...]**. Belo Horizonte: UFMG, 2014. p. 1-21. Disponível em: <http://hdl.handle.net/123456789/2598>. Acesso em: 06 de julho de 2023.

MOREIRO-GONZÁLEZ, J. A. **Linguagens documentárias e vocabulários semânticos para a web**: elementos conceituais. Salvador: UDFBA, 2011.

PINHEIRO, L. V. R. Informação: esse obscuro objeto da ciência da informação **Revista Morpheus**: Revista Eletrônica em Ciências Humanas: conhecimento e sociedade, Rio de Janeiro, v. 3, n. 4, p. 01-11, 2004.

PINHEIRO, L. V. R. Mutações na ciência da informação e reflexos nas mandalas interdisciplinares. **Informação & Sociedade: Estudos**, João Pessoa, v. 28, n. 3, 2018. DOI: 10.22478/ufpb.1809-4783.2018v28n3.43317. Acesso em: 24 jun. 2023.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesouro Brasileiro de Ciência da Informação**. [PDF]. Rio de Janeiro, Brasília: IBICT, 2014.

SCHÄRER, M. Things + ideas + musealization = heritage a museological approach. **Revista Museologia e Patrimônio**, Rio de Janeiro, v. 2, n. 1, p. 85-89. 2009.

SILVA, M. dos R. DA.; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transformação**, Campinas, SP, v. 12, n. 2, p. 133-161, 2004.

SCHWING, T.; MCCUTCHEON, S.; MAURER, M. B. Uniqueness matters: author-supplied keywords and LCSH in the library catalog. **Cataloging & Classification Quarterly**, [S. l.], v. 50, n. 8, p. 903-928, 2012.

SMIT, J. W.; KOBASHI, N. Y. **Como elaborar vocabulário controlado para aplicação em arquivos**. São Paulo: Arquivo do Estado, Imprensa Oficial, 2003. (Projeto como fazer, 10).

STRADER, C. R. Author-assigned keywords versus library of congress subject headings: implications for the cataloging of electronic theses and dissertations. **Library Resources & Technical Services**, [S. l.], v. 53, n. 4, p. 243-50, 2009.

Disponível em:

<https://www.journals.ala.org/index.php/Irts/article/viewFile/5183/6291/>. Acesso em 06 jul. 2023.

SUAIDEN, E. J. Apresentação. In: PINHEIRO, L. V. R.; FERREZ, H. D. **Tesouro Brasileiro de Ciência da Informação**. Rio de Janeiro, Brasília: IBICT, 2014.

TERRA, A. L.; AGUSTÍN LACRUZ, C; BERNARDES, O.; FUJITA, M. S. L. Subject-access metadata on ETD supplied by authors: A case study about keywords, titles and abstracts in a Brazilian academic repository. **Journal of Academic Librarianship**, v. 47, n. 1, p. 102268, 2021.

VIGNOLI, R. G.; SOUTO, D. V. B.; CERVANTES, B. M. N. Sistemas de organização do conhecimento com foco em ontologias e taxonomias. **Informação & Sociedade: Estudos**, João Pessoa, v. 23, n. 2, p. 59-72, maio/ago. 2013.

WU, Y. Enriching a thesaurus as a better question-answering tool and information retrieval aid. **Journal of Information Science**, United States, v. 44, n. 4, p. 512-525, 2017.

ASSIGNMENT OF KEYWORDS FOR COMPLETE WORKS AT THE EVENT: ANALYSIS OF VOCABULARY CONTROL

Objective: In order to verify the indexing process in the attribution of keywords in event publications in Information Science, an investigation was carried out on vocabulary control using the Brazilian Thesaurus in Information Science (TBCI) by the authors on the Portal of the Meeting of the National Association for Research in Information Science (ENANCIB). **Methodology:** The study had the ENANCIB Annals as a research universe and, as a sample, publications of the complete work modality of the Annals of 2015 were

analyzed, the year in which vocabulary control began to be recommended until 2018. It was performed an analysis of the keywords assigned by the authors, after extraction and mining with the Keyword Extraction Software (SEP), created for this purpose. **Results:** 1,075 articles from the 11 Working Groups of the event were retrieved, with a total of 4,437 keywords examined. The main results demonstrate that 60.3% of the keywords retrieved in the articles are present as controlled terms in the Thesaurus. Working Group 4 obtained the best performance, in the analysis carried out, and Work Group 10, the most incipient result. **Conclusions:** The collected data indicate that changes are necessary for the ENANCIB Annals to have a more comprehensive vocabulary control. It is recommended to create tutorials for authors, with guidelines and benefits on using the Thesaurus for vocabulary control.

Descriptors: Authors supplied indexing. Vocabulary control. Thesaurus. Keywords.

ASIGNACIÓN DE PALABRAS CLAVE EN TRABAJOS COMPLETOS DEL EVENTO: ANÁLISIS DE CONTROL DE VOCABULARIO.

RESUMEN

Objetivo: Para verificar el proceso de representación en la atribución de palabras clave en publicaciones de eventos en Ciencias de la Información, se realizó una investigación sobre el control de vocabulario utilizando el Tesoro Brasileño en Ciencias de la Información (TBCI) por los autores en el Portal del Encuentro de la Asociación Nacional de Investigación en Ciencias de la Información (ENANCIB). **Metodología:** El estudio tuvo como universo de investigación los Anales del ENANCIB y, como muestra, se analizaron publicaciones de la modalidad de trabajo completo de los Anales del año 2015, año en el que se comenzó a recomendar el control de vocabulario hasta el 2018. Palabras clave asignadas por los autores, previa extracción y minado con el Software de Extracción de Palabras Clave (SEP), creado para tal fin. **Resultados:** Se recuperaron 1.075 artículos de los 11 Grupos de Trabajo del evento, con un total de 4.437 palabras clave examinadas. Los principales resultados demuestran que el 60,3% de las palabras clave recuperadas en los artículos están presentes como términos controlados en el Tesoro. El Grupo de Trabajo 4 obtuvo el mejor desempeño, en el análisis realizado, y el Grupo de Trabajo 10, el resultado más incipiente. **Conclusiones:** Los datos recopilados indican que son necesarios cambios para que los Anales del ENANCIB tengan un control de vocabulario más integral. Se recomienda crear tutoriales para autores, con pautas y beneficios sobre el uso del Tesoro para el control de vocabulario.

Descriptor: Indización por autores. Control de vocabulario. Tesoro. Palabras clave.

Recebido em: 14.01.2023

Aceito em: 14.07.2023