DOI: 10.5433/1981-8920.2020v25n3p332

ANÁLISE DE PALAVRAS-CHAVE DA PRODUÇÃO CIENTÍFICA DE PESQUISADORES: O AUTOR COMO INDEXADOR

ANALYSIS OF KEYWORDS OF SCIENTIFIC PRODUCTION OF RESEARCHERS: THE AUTHOR AS AN INDEXER

Mariângela Spotti Lopes Fujita[a]
Roberta Cristina Dal'Evedove Tartarotti[b]

RESUMO

Introdução: Com o advento da era digital e web, a palavra-chave tornou-se um produto de representação essencial em sistemas de armazenamento e recuperação da informação de acesso aberto. É utilizada nas funções de extração para fins de identificação científica de autores, análise bibliométrica, indicadores de impacto científico, desenvolvimento de vocabulários controlados e outros sistemas de organização do conhecimento. A atribuição de palavras-chave pelo autor, em publicações científicas, é uma prática de representação do conteúdo, realizada durante o preenchimento de metadados. Essa atribuição passa por análise de assunto personalizada e depende do vocabulário de especialidade do autor que, de modo geral, não tem orientação sobre padronização e controle de vocabulário. Por outro lado, tais metadados de assunto que recebem a palavra-chave não passam por validação profissional. **Objetivo:** Análise de palavras-chave atribuídas por pesquisadores para submissão de artigos de periódicos indexados na Scopus e no Portal Docentes Unesp, quanto à padronização e controle de vocabulário para diferentes funções em sistemas de armazenagem e recuperação da informação. **Metodologia:** Para tanto, foi feita pesquisa exploratória com estudo de observação e análise de palavras-chave atribuídas por pesquisadores do Portal Docentes Unesp, a partir dos dados do Currículo Lattes do CNPq, comparadas com palavras-chave atribuídas aos artigos de periódicos. **Resultados:** Os resultados demonstram ausência de padronização nas palavras-chave dos artigos de periódicos dos pesquisadores, cadastradas no Currículo Lattes, tanto em nível sintático quanto semântico. Quanto à avaliação da indexação, observam-se baixos índices de consistência, quando comparadas aos artigos originais, tanto no índice rígido quanto no relaxado/flexível. **Conclusões:** Verifica-se a necessidade de elaboração de uma política de organização e representação da informação que forneça diretrizes aos pesquisadores, quanto à atribuição de palavras-chave, visando a uma maior

[a] Doutora em Ciências da Comunicação pela Universidade São Paulo. Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP). E-mail: mariangela.fujita@unesp.br.

[b] Doutora em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP). Bibliotecária de Referência no Centro de Recursos de Aprendizagem da Biblioteca Central César Lattes (BCCL) da Unicamp. E-mail: roberta_tartarotti@yahoo.com.br.

padronização e consistência, tanto na representação quanto na recuperação de sua produção científica.

Descritores: Palavra-chave. Indexação. Produção científica. Controle de vocabulário. Vocabulário controlado. Currículo Lattes.

1 INTRODUÇÃO

A atribuição de palavras-chave pelo autor, em vários sistemas de informação, não é recente, porém, imprecisa quanto ao uso. Smiraglia (2013) refere-se ao surgimento da palavra-chave nos primeiros dias da indexação automatizada, quando foi descoberto que a utilização de termos de títulos e resumos de textos completos tinha função “chave”, na busca simples ou avançada, com o uso de lógica booleana.

O emprego da palavra-chave vai além da busca e passa a ter aplicação na indexação, recuperação de informações, marcação social, extração de palavras-chave, bibliometria e desenvolvimento de tesauros e outros sistemas de organização do conhecimento (LU; LI; ZHIFENG; CHENG, 2019, p. 415).

A relevância de palavras-chave para tais aplicações reside no fato de que é o próprio autor quem garante a representatividade “chave” de seus textos. Isso é compreensível, porque o autor produziu o texto, é o especialista do tema tratado e tem domínio do conteúdo e do vocabulário utilizado. Portanto, é ideal que ele próprio atribua palavras-chave que representem o texto por ele produzido.

Garcia, Gattaz e Gattaz (2019, p. 3) consideram que palavras-chave são “[...] fundamentais para que os textos sejam capturados pelos mecanismos de buscas e alcancem seus possíveis leitores” e, em consequência, favoreçam a comunicação científica, com o uso de citações que serão aferidas por sistemas de informação.

Em investigação sobre padrões potenciais de palavras-chave selecionadas por autores em artigos científicos, Lu, Li, Zhifeng e Cheng (2019) realizaram análise de palavras-chave, com a finalidade de identificar a função de cada termo e contribuir para a construção de rede semântica de palavras-chave importante para a organização do conhecimento, bem como para tarefas

bibliométricas. Os resultados analisados descreveram irregularidade das palavras-chave, na perspectiva da função do termo, confirmaram que as palavras-chave sobre “tema de pesquisa” e “método de pesquisa” são as mais frequentes em artigos científicos, tendo verificado três padrões de palavras-chave que dependeram da quantidade de palavras-chave, como, por exemplo, “tópico de pesquisa”, em que as palavras-chave apareciam mais nas três primeiras posições, e “métodos de pesquisa”, quando as palavras-chave apareciam com maior probabilidade nas duas últimas posições da lista. É interessante observar que, ao final, os autores produziram lista de palavras-chave organizadas quanto à função que lhes foram atribuídas.

Por outro lado, Han, Harrington, Black e Kudeki (2016) examinaram as possibilidades de comparar palavras-chave atribuídas pelos autores de teses e dissertações eletrônicas com o vocabulário controlado da *Library of Congress Subject Headings* (LCSH) e verificaram, também, se elas poderiam se alinhar melhor com outros vocabulários mais específicos. O resultado obtido, segundo os autores, não foi muito bom, porque muitas palavras-chave somente são utilizadas em uma tese ou dissertação. Observaram, contudo, que o uso combinado de outros vocabulários controlados específicos com a LCSH produz um melhor resultado de compatibilização com palavras-chave.

Em editorial da *Knowledge Organization,* Smiraglia (2013) revelou sua decepção, ao descobrir como havia pouca correlação entre a indexação formal de um pequeno conjunto de artigos do periódico e, especialmente, por ver como havia pouca correspondência entre as palavras-chave reais que apareciam nos textos publicados e qualquer uma das indexações fornecidas pela *Web of Science* ou LISTA. Isso também demonstra a incompatibilidade entre palavras-chave e descritores de vocabulários controlados, muitas vezes provocada por prováveis causas, como desatualização de vocabulários controlados ou falta de especificidade de descritores.

Em periódicos que publicam artigos, Fujita, Agustín-Lacruz e Terra, (2018a; 2018b), analisaram as diretrizes aos autores, para submissão de originais, sobre a redação do título, resumo e palavras-chave de seus artigos, em uma amostra representativa dos periódicos de Biblioteconomia e Ciência da

Informação (LIS) e Ciência da Comunicação (CS) indexados no Journal Citation Relatórios (JCR). As conclusões alertam para indicações e critérios gerais idênticos nas diretrizes de vários periódicos gerenciados por editoras comerciais, as quais fornecem plataformas de acesso aos periódicos e poucas instruções sobre títulos, resumos e palavras-chave.

Em decorrência disso, o autor, quando atribui palavras-chaves, preenche um metadado, sem orientação ou auxílio profissional, que será preservado quanto à padronizações ou qualquer proposta de controle de vocabulário, após preenchimento. O autor talvez não preveja outras finalidades de uso por sistemas de informação com os quais não necessariamente tenha ligação direta, mas sim indireta. Prova disso são os casos de artigos de periódicos posteriormente indexados e representados por descritores de vocabulários controlados ou quando se cadastra em plataformas digitais e registra sua produção científica para divulgação em redes sociais acadêmicas, cujo conjunto de dados é posteriormente analisado por outras plataformas. Significa que o autor, ao atribuir palavras-chave torna-se um indexador e precisa pensar em outras funções além da representação do conteúdo do texto e adotar padrões que sejam compatíveis com diferentes empregos futuros.

Com o objetivo de analisar palavras-chave atribuídas por pesquisadores para submissão de artigos de periódicos indexados na Scopus e no Portal Docentes Unesp, quanto à padronização e controle de vocabulário para diferentes funções em sistemas de armazenagem e recuperação da informação, será implementada pesquisa exploratória com estudo de observação e análise de palavras-chave atribuídas pelos pesquisadores do Portal Docentes Unesp, sistema de informação que realiza a gestão e a divulgação científica dos docentes da Unesp, a partir dos dados do currículo Lattes do CNPq.

2 FUNDAMENTAÇÃO TEÓRICA

Os produtos da representação documentária mais visíveis e utilizados para a recuperação da informação em bases de dados digitais são o título, resumo e palavras-chave. Todos são produtos documentários cuja característica em comum é o processo de condensação do conteúdo documentário. Segundo

Gonçalves (2008), essa característica favorece a comunicação científica, mesmo com o aumento progressivo da quantidade de publicações, porque possibilita a leitura do documento por meio de sua representação condensada.

A origem da palavra-chave não tem precisão exata, mas podemos afirmar que está ligada aos interesses de recuperação da informação, desde as primeiras iniciativas de indexação manual por palavras do título, feitas por Crestadoro, na compilação do catálogo da Biblioteca Pública de Manchester, no século 19 (FOSKETT, 1973). Porém, é na indexação automática que a palavra-chave tem sua função fundamental para a recuperação da informação, através das propostas preconizadas por Hans Peter Luhn (FOSKETT, 1973). Sua origem e funcionalidade podem, também, ser explicadas pela proposta do Unitermo de Mortimer Taube, com o sistema *Coordinate Indexing*, criado por ele e Alberto E. Thompson (GULL, 1987). De acordo com Chu (2010), unitermos podem ser vistos hoje como palavras-chave, porque ambos são derivados de documentos originais, sem nenhum esforço de controle de vocabulário.

O uso de palavra-chave está relacionado, assim, à expressão de uma palavra significativa que seja “chave” na representação de um texto ou recurso informacional, para que seja acessado e recuperado em diversas outras fontes de informação. Assim, essa “chave”, se sistematicamente extraída ou atribuída, pode abrir portas para outras possibilidades de divulgação e dar importância ao recurso informacional.

O conceito de palavra-chave está condicionado à representação do significado de um determinado conteúdo verbal ou não verbal e tem diversas finalidades, mas, essencialmente, é utilizado para identificação de ideias e temas importantes. Podemos considerar que a palavra-chave tem expressiva quantidade de finalidades diversas, as quais incluem estudos bibliométricos, indexação, recuperação etc.

A indexação pode ser realizada, segundo Lancaster (2004), por extração ou atribuição. Para construção dos índices KWIC (*Key-Word In Context*), Luhn desenvolveu software de indexação automática, a partir de palavras-chave extraídas dos títulos de publicações (FOSKETT, 1973; CHU, 2010). A indexação com palavras-chave atribuídas pelo autor ou usuário, por outro lado, está

fortemente associada a uma representação semântica do que consideram significativo.

No âmbito da indexação social por *tagging* ou marcação, Moulaison (2008) enfatiza que existem diferenças entre *endo-tagging*, metadados fornecidos pelo autor e *exo-tagging*, metadados fornecidos pelo usuário. De fato, a marcação social pelo usuário trouxe outra função importante à palavra-chave, embora se necessite de mais análise sobre o tema.

A palavra-chave é, de fato, a representação documentária com maior grau de condensação do que o título e o resumo. Sua utilidade na comunicação científica se faz presente em muitos tipos de publicação. Com a internet, a web e a filosofia do acesso aberto, a submissão de artigos em periódicos, trabalhos em eventos, projetos de pesquisa em agências de fomento, teses e dissertações em repositórios exigem de seus autores, além de outros metadados, a atribuição de palavras-chave capazes de representar o significado mais importante do conteúdo. Essa atribuição, embora aparentemente simples, exige padronização, equivalência e consistência entre as demais publicações do mesmo autor, bem como para as diferentes finalidades, tendo em vista que as palavras-chave serão posteriormente adotadas para a definição de perfil científico, em diversas fontes de informação.

O produto da representação documentária, seja palavra-chave, seja descritor, autor, título ou outros, está abrigado em metadado correspondente, para que seja posteriormente recuperado. Metadado refere-se “[...] às descrições especificamente criadas para representar informação digital acessível na internet” (CHU, 2010, p. 41).

Apesar de acreditarmos que sabemos o que é metadado, porque é um recurso de padronização importante usualmente conhecido, é necessário distinguirmos, em seu significado, duas definições que fazem a diferença no quesito de representação do conteúdo. Chu (2010) esclarece que o metadado é definido de dois modos: “[...] em sentido específico, implica descrições fornecidas para informações em rede e recursos digitais, seguindo um padrão ou estrutura que é especificamente criado para este fim” (CHU, 2010, p. 41), como, por exemplo, o *Dublin Core*. Porém, de modo geral, “[...] tem uma cobertura mais

ampla, incluindo dados de catalogação e indexação criados para qualquer tipo de documento por meio do uso de métodos tradicionais para descrever e organizar informações" (CHU, 2010, p. 41).

Em específico, metadados de assuntos são subdivididos em dois tipos distintos, conforme preenchimento: metadados de vocabulário controlado extraídos de listas de termos e metadados de texto livre. No caso de bibliotecas digitais ou repositórios que agregam metadados de diferentes fontes, os problemas de consistência devem ser avaliados com métodos de pesquisa intrusivos, tais como análise de conteúdo, análise de log de transações ou observação, entre outros (ZAVALINA, 2011b). Recomenda-se ainda, consulta às orientações da IFLA acerca da definição de um modelo de referência conceitual para análise de metadados (INTERNATIONAL, 2017)

No que se refere aos metadados fornecidos pelo autor, estudos sobre o processo de atribuição de palavras-chave foram realizados por Fujita (2004), Gil Leiva e Alonso Arroyo (2007), Joorabchi e Mahdi (2012), Kun e Kipp (2014), Li (2018), Vanyushkin e Graschenko (2020).

Com proposta de explicitar o processo de atribuição de palavras-chave pelo autor, em artigos de periódicos da área de Educação Especial, Fujita (2004) descreve orientação aos autores com base em metodologia de indexação. Recomenda a aplicação de estratégia de análise de assunto com localização do tema principal na estrutura textual do artigo e compatibilização das palavras-chave, com a utilização de vocabulário controlado da área de Educação.

Em análise comparada de palavras-chave fornecidas pelos autores de artigos científicos e os descritores atribuídos pelos serviços de indexação das bases de dados bibliográficas INSPEC, CAB, ISTA e LISA, Gil-Leiva e Alonso-Arroyo (2007) constataram que 46% das palavras-chave atribuídas por autores têm importante compatibilidade com os descritores atribuídos pelas bases de dados, sendo que 25% são idênticas. Os autores do artigo também verificaram que três das quatro bases de dados possuem política de indexação distintas para seus indexadores.

O artigo de Joorabchi e Mahdi (2012) focaliza método experimental de anotação de frase-chave automática que utiliza a Wikipedia, como dicionário de

sinônimos para seleção de frases candidatas, a partir do conteúdo dos documentos. Os resultados obtidos foram avaliados em termos de interconsistência com anotadores humanos, observando-se que estão no mesmo nível alcançado por humanos.

Na investigação sobre a eficácia da recuperação de tags colaborativas e palavras-chave do autor, o estudo comparado elaborado por Kun e Kipp (2014) avaliou o desempenho de recuperação em três experimentos controlados: o primeiro, quando apenas o título e o resumo estão disponíveis; o segundo, quando o texto completo está disponível; e o terceiro, comparando a recuperação de tags e palavras-chave do autor apenas no resumo. Os resultados obtidos demonstraram que as palavras-chave do autor são mais vantajosas para aumentar a recuperação e que os sistemas de recuperação devem incorporar tags e palavras-chave do autor.

Munan (2018) considera que as palavras-chave do autor desempenham função fundamental na análise bibliométrica tradicional, no que concerne, especialmente, à análise de copalavras, consulta de informações e captura de termos tópicos que irão definir perfis científicos dos pesquisadores e de periódicos. Por isso, o estudo investiga métodos de ponderação de palavras-chave de autor e explora a questão de diferenciação de funções para a atribuição de palavras-chave do autor, por meio de estrutura analítica baseada na diferenciação de funções e Técnica para Ordem de Preferência por Similaridade à Solução Ideal.

Em experimento sobre avaliação de algoritmos de extração automática de palavras-chave, Vanyushkin e Graschenko (2020) concluíram que existe uma dependência com relação às propriedades das coleções de teste, a qual torna muito difícil comparar diferentes algoritmos; além disso, é difícil prever a eficácia de diferentes sistemas, a fim de resolver problemas do mundo real de processamento de linguagem natural. Entretanto, ponderam que, até o momento, não apareceu uma técnica consistente para detectar palavras-chave em textos, a despeito da grande quantidade de trabalhos especializados e interdisciplinares. Relatam que outros experimentos confirmam que a atribuição de palavras-chave é feita de forma intuitiva e personalizada, por isso, os esforços

desses pesquisadores estão voltados para o desenvolvimento de algoritmos de extração automática de palavras-chave, com base na aprendizagem e com o uso de recursos linguísticos.

A avaliação ou a análise do processo de atribuição de palavras-chave pelo autor como indexador foram realizadas por estudos de Kipp (2009), Névéol, Doǧan e Zhiong (2010). Conforme esses pesquisadores, o autor é o especialista e precisa conhecer seu campo de pesquisa, o domínio científico no qual está inserido.

Em artigo sobre a atribuição de palavras-chave por autores em artigos de periódicos biomédicos, Névéol, Doǧan e Zhiong (2010) apresentam um estudo comparativo de palavras-chave de autores e termos de indexação MeSH, atribuídos por indexadores MEDLINE a artigos de acesso aberto, cujos resultados mostram que as palavras-chave do autor estão cada vez mais disponíveis em artigos biomédicos e que cerca de 60% das palavras-chave do autor são compatíveis com termos de indexação.

Os autores assinalam que há diferenças significativas, em termos de forma e perspectiva, entre a atribuição de palavras-chave pelo autor a um artigo e a atribuição de termos de indexação pelo indexador, embora possam parecer semelhantes. As principais diferenças são: a escolha de palavras-chave sem referência a um vocabulário controlado, feita pelo autor, enquanto os indexadores são treinados para selecionar termos de indexação conforme procedimentos específicos de uma política de indexação; os autores não têm um objetivo profissional na seleção de palavras-chave que representem o que consideram importante para descrever o conteúdo de seu próprio artigo, ao passo que os indexadores precisam levar em conta o conteúdo do artigo, dentro do âmbito maior da coleção.

Nessa mesma linha de investigação, Kipp (2009) verificou que há diferenças de resultados de indexação entre três grupos distintos – usuários, autores e intermediários –, observadas a partir de análise de artigos de periódicos marcados no CiteULike. A mais significativa, entretanto, é o fato de que catalogadores e indexadores que atuam como intermediários classificam e descrevem os documentos de acordo com padrões estritos de consistência do

termo e adesão a um conjunto de decisões de política estabelecidas em padrões de catalogação, sistemas de classificação e vocabulários controlados.

Todavia, os autores ponderam que vocabulários controlados precisam de treinamento para usá-los e são caros para aplicar. A marcação colaborativa permite que usuários participem da classificação de artigos de periódicos, e essa é uma habilidade inata comum a todos os humanos. O estudo realizado também demonstra que algumas diferenças entre marcação de usuário e indexador são distinções de palavras que poderiam ser superadas, por meio de truncamento ou lematização, além de existir igualmente outros vocabulários que diferem do vocabulário profissional, os quais poderiam ser aplicados durante a prática de atribuição de palavras-chave.

Estudos de análise de palavras-chave foram feitos por Gonçalves (2008), Jennifer e Muthuhumaravel (2019), Peset *et al.* (2020), no sentido de propor metodologias de análise diferenciadas que trouxeram resultados inovadores para a padronização de palavras-chave.

Ao avaliar o potencial das palavras-chave como complemento de resumos de artigos científicos das disciplinas de Antropologia, Ciência Política e Sociologia, Gonçalves (2008) observou a falta de padronização em palavras-chave fornecidas pelos autores e que estes podem complementar o resumo, assim como atuar como indicadores de relações conceituais. Entretanto, conclui que, apesar da importância dos resumos e palavras-chave como elementos de representação da informação, suas funções estratégicas são desconhecidas e pouco exploradas, nas Ciências Sociais.

Jennifer e Muthukumaravel (2019) apresentam estudo experimental em sistemas de recuperação da informação, cuja ideia básica é a busca por palavras-chave em documentos com grande espaço de busca, considerando-se que cada texto tenha tamanho diferente. Asseveram que, ao se reduzir o tempo de pesquisa por palavras-chave, poderão reduzir o espaço de pesquisa e, dessa forma, agilizar a busca e a recuperação da informação. Para isso, constroem um método de redução da área de pesquisa, com ajuda da indexação, o qual usa o método de lematização e conhecimento de palavras irrelevantes.

Em estudo bibliométrico sobre análise de sobrevivência de palavras-

chave do autor, na área de Ciência da Informação, Peset *et al.* (2020) adaptaram um método estatístico de análise de séries numéricas discretas às palavras-chave que aparecem em artigos, com o objetivo de detectar as novas palavras-chave de autores, no período de um ano, para quantificar as probabilidades de sobrevivência por 10 anos, em função do impacto dos periódicos onde aparecem. Os resultados obtidos demonstram que o tempo médio de sobrevivência dessas palavras-chave é de três anos, e um pouco maior, em casos de artigos publicados no segundo quartil da área; muitas novas palavras-chave que aparecem na área de Ciência da Informação são efêmeras e mais da metade nunca é usada novamente. Os autores observaram ainda que muitas palavras-chave comumente empregadas em Ciência da Informação são de áreas interdisciplinares.

No que se refere à recuperação da informação, a qualidade da representação contida no metadado de assunto assume fundamental importância, em pesquisas de avaliação efetuadas por Newman, Hagedorn e Smyth (2007), Zavalina (2011a), Yang (2016), Hanrath e Radio (2017), Balatsoukas, Rousidis e Garoufallou (2018).

Hanrath e Radio (2017) consideram necessário aumentar a atenção à descrição de assunto em metadados, a fim de melhorar a visibilidade da pesquisa de conteúdos, com base em resultados de investigação do comportamento de pesquisa de usuários em relação a assuntos, com base na aplicação de um vocabulário controlado em um repositório institucional. Metadados de baixa qualidade resultam em impacto negativo, na recuperação e no gerenciamento de repositórios de dados de pesquisa, conforme pesquisa de Balatsoukas, Rousidis e Garoufallou (2018) sobre análise de metadados de assuntos do Dublin Core, cujos resultados mostraram problemas de qualidade relacionados à falta de controle de vocabulário e padronização, como o uso de formas singulares e plurais, adjetivos e sinônimos. Metadados de assunto uniformes, consistentes e enriquecidos, segundo Newman, Hagedorn e Smyth (2007), permitem que os usuários descubram o material com mais facilidade, naveguem na coleção e limitem os resultados da pesquisa de palavras-chave por assunto. Para isso, realizaram proposta de enriquecimento de metadados de

assunto, com a utilização de técnicas de modelagem manual e estatística.

O estudo de avaliação comparativa de metadados de assunto em coleção de texto livre, feito por Zavalina (2011b), observou que as melhores práticas incluem, no elemento de descrição, variedade de informações específicas do assunto tópico, cobertura geográfica e temporal e tipos/gêneros de objetos em uma coleção digital, além de outras informações não específicas. Em outra pesquisa, Zavalina (2011a) analisou o papel dos metadados em nível de coleção, na recuperação de informações em agregações digitais, com base em consultas de pesquisa do usuário derivadas de logs de transações, combinada com estudo de usuário composto de entrevista e observações de usuários interagindo com o sistema. Os usuários têm preferência por visualizar registros completos de metadados estruturados em nível de coleção, os quais incluem metadados de assuntos. Resultados de investigação de Yang (2016), examinando metadados de itens digitais e palavras-chave de motores de busca da internet, para verificar quais elementos de metadados facilitam a descoberta de coleções digitais, indicam que os elementos de metadados título, descrição e assunto são eficazes para aumentar a capacidade de descoberta.

Nesse sentido, ganham relevância os estudos sobre a padronização e a normalização dos termos de vocabulários controlados, visando à recuperação e visibilidade da produção científica dos pesquisadores em bases de dados (WILLIS; LEE, 2013; FROTA, 2014; SANTOS; CERVANTES, 2015; HIDER, 2018; GOLUB; TYRKKÖ; HANSSON; AHLSTRÖM, 2020).

Com o objetivo de análise da contribuição da estrutura do tesauro para o processo de indexação, os pesquisadores Willis e Lee (2013) apresentam uma metodologia de análise e modelagem do processo de indexação, com base em algoritmo de caminhada aleatória ponderada. Os resultados obtidos são avaliados no contexto da indexação automática de assuntos, usando quatro coleções de documentos pré-indexados com quatro tesauros diferentes: AGROVOC, taxonomia de física de alta energia (HEP), National Agricultural Library Thesaurus (NALT) e Medical Subject Headings (MeSH). Em todos os casos, a metodologia utilizada melhorou o desempenho da indexação automática e os autores ponderam que a indexação de assuntos é, em parte, um

processo de navegação e que o uso do vocabulário e sua estrutura contribui e influencia o processo de indexação e, consequentemente, a recuperação.

A análise do Tesauro da Corte Interamericana de Direitos Humanos, realizada por Frota (2014), teve o objetivo de apontar elementos que cooperem para o aperfeiçoamento do Tesauro da Corte. Foram constatadas, na análise dos resultados, incoerência ou incompletude em relação à literatura adotada para coleta de termos da área de Direitos Humanos, formação inadequada de termos gerais e específicos para representar as dimensões dos Direitos Humanos, incoerência entre conceitos e emprego de termos em desuso. O autor recomenda a revisão do Tesauro, a fim de melhorar os relacionamentos entre termos e a adequação à linguagem especializada da área.

Santos e Cervantes (2015) utilizam ferramenta complementar ao Sistema Eletrônico de Editoração de Revistas (SEER) do IBICT, para analisar as condições de funcionamento de um controle de vocabulário de palavras-chave atribuídas em periódicos científicos eletrônicos. Para isso, os autores promoveram o controle de vocabulário da palavras-chave e a compatibilização entre os termos dos vocabulários controlados e as palavras-chave, com o emprego da ferramenta. Os resultados obtidos permitiram observar a diversidade de sintaxe das palavras-chave atribuídas e, por outro lado, levaram a construir um índice de termos que poderá auxiliar os autores na atribuição de palavras-chaves.

Resultados de estudos anteriores revelaram que as consultas de tópicos de assuntos perderiam cerca de um quarto dos acessos, em média, caso não incorporassem a indexação de assuntos com vocabulários controlados. Em pesquisa de replicação da metodologia anteriormente empregada por Gross *et al.* (2015), Hider (2018) investigou o nível de perdas em consultas de tópicos de assuntos, sem a incorporação da indexação de assuntos baseadas em vocabulários controlados. Para isso, Hider (2018) executou consultas de assuntos em três bancos de dados bibliográficos, ERIC, PsynINFO e Socindex, de sorte a investigar se o nível de perda poderia ser generalizável. Com variações de perdas entre os bancos de dados bibliográficos avaliados, percebe-se que é necessário manter a indexação atribuída nessas bases de dados.

Em pesquisa recente sobre indexação de assuntos em Humanidades,

Golub, Tyrkkö, Hansson e Ahlström (2020) realizaram comparação entre um Repositório Universitário da Suécia e a Scopus, uma bibliografia internacional, com o objetivo de apresentar a situação atual do uso de termos de índice de assuntos em artigos de periódicos de Ciências Humanas, de modo a identificar aprimoramentos necessários nesse contexto. Os resultados obtidos revelaram que nenhum vocabulário controlado para qualquer disciplina de Ciências Humanas é usado no Repositório Universitário e na Scopus.

Na perspectiva da padronização e normalização de palavras-chave, estudo de Toepfer e Seifert (2017) sobre a proposta de combinação de palavras-chave com vocabulários controlados, esta poderá ser avaliada de forma positiva, durante a atribuição de palavras-chave pelos autores. Nessa proposta, os experimentos efetuados com sistemas de indexação automática que utilizam componentes lexicais, aplicando correspondência de dicionário, classificação e classificação binária para avaliar títulos econômicos e palavras-chave do autor, mostram que as estratégias de fusão, combinando uma abordagem de relevância binária e um sistema baseado em dicionário de sinônimos superam todas as outras estratégias, no conjunto de dados testado. Essa proposta é particularmente estratégica, para que sistemas que adotam palavras-chave e linguagem natural tenham disponibilização de controle de vocabulário, a fim de que seja possível consistência entre representação na atribuição de assuntos por autores e recuperação por usuários.

O tema de combinação de controle de vocabulário com linguagem natural traz à tona o debate instalado desde o aparecimento das primeiras bases de dados bibliográficas, com o emprego de palavras-chave extraídas dos títulos de artigos de periódicos, a partir da metade do século XX, e que passaram a coexistir com os sistemas de classificação, os vocabulários controlados predominantes desde o século XIX. A evolução desse debate sobre controle de vocabulários é explicada por Rowley (1994) e Chu (2010), no que elas denominaram as quatro eras do debate, sintetizado no Quadro 1, abaixo:

Quadro 1 – As quatro eras do debate

ERA UM	Vocabulário controlado.
ERA DOIS	Comparações entre linguagem controlada e natural – Qual linguagem é melhor?

	Linguagem natural pode desempenhar tão bem quanto vocabulário controlado, mas outros fatores, tais como o número de pontos de acesso, também são significativos; Estudos experimentais importantes.
ERA TRÊS	Muitos estudos de caso de generalização limitada. Buscas em bases de dados online são consideradas. Melhor desempenho pela combinação de linguagem controlada e natural. Reafirmação de que o número de pontos de acesso tem efeito significativo. Distinções entre texto completo e bases de dados bibliográficas.
ERA QUATRO	Novos avanços em sistemas baseados em usuários, incluindo OPACs. Renascimento do valor de vocabulários controlados, no contexto de interfaces amigáveis ao usuário, e o desenvolvimento de bases de conhecimento.

Fonte: Rowley (1994, p. 110)

Outros estudos sobre linguagem natural *versus* linguagem controlada analisam o comportamento de recuperação por palavras-chave da linguagem natural, comparado a descritores de uma linguagem controlada (GROSS; TAYLOR, 2005; GROSS; TAYLOR; JOUDREY, 2015; ZHANG *et al.*, 2016; BOGERS; PETRAS, 2015; KIYOTA *et al*., 2008; MAURER; SHAKERI, 2013; HORN, 2002; SCHWING; MCCUTHEON; MAURER, 2012).

Santos e Corrêa (2017) procederam à avaliação da política de indexação da Base de Dados Referencial de Periódicos em Ciência da Informação (BRAPCI), a qual armazena, preserva e divulga a memória científica brasileira da área de Ciência da Informação. A partir da identificação e análise dos problemas, os autores recomendam a elaboração de política de indexação que proporcione o aprimoramento das práticas e a adoção de vocabulários controlados capazes de auxiliar a descrição dos documentos, a fim de se obter mais precisão no processo de busca e recuperação. Nesse caso, os documentos armazenados são coletados e migrados de fontes diversas, e parte dos documentos não tem atribuição de palavras-chave pelo autor, o que torna ainda mais necessária a política de indexação destinada a guiar os profissionais que realizarão a indexação.

3 PROCEDIMENTOS METODOLÓGICOS

Visando à padronização da representação que favoreça a elaboração de política de indexação para orientação, quanto ao aprimoramento do uso de

palavras-chave para diferentes funções, os procedimentos metodológicos do trabalho pautaram-se em duas principais frentes: a) revisão de literatura sobre a atribuição de palavras-chave do autor como indexador, no contexto científico; b) pesquisa exploratória com estudo de observação e análise de palavras-chave atribuídas pelos pesquisadores do Portal Docentes Unesp, sistema de informação que realiza a gestão e divulgação científica dos docentes da Unesp, com base nos dados do Currículo Lattes do CNPq.

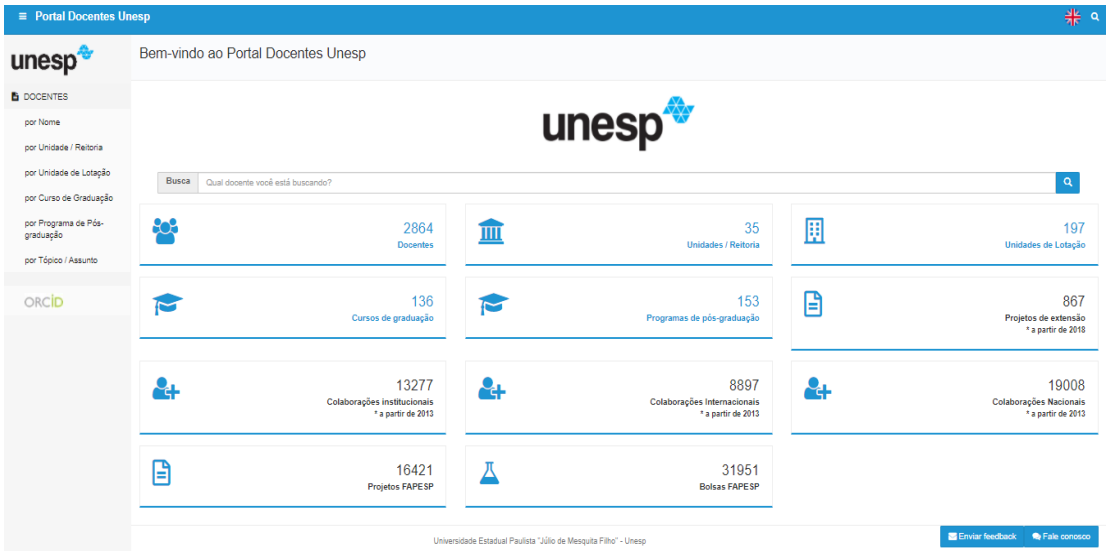
Lançado em junho/2019, o Portal Docentes Unesp[1] é uma plataforma que reúne informações sobre os docentes da Universidade, vinculados a todas as áreas do conhecimento, desenvolvida por meio de parceria entre as Pró-Reitorias, Coordenadoria Geral de Bibliotecas (CGB), Coordenadoria de Tecnologia da Informação (CTINF) e Assessoria de Relações Externas (Arex). Possibilita buscas em português e em inglês e utiliza, como fonte de dados, 13 diferentes fontes: ORCID, Biblioteca Virtual da Fapesp, Capes, Plataforma Lattes, *Google Scholar*, *Web of Science* e Scopus (fontes externas); Repositório Institucional da Unesp e sistemas institucionais da Unesp (SISRH, SISGRAD, SISPG, SISPROPE e SISPROEX) (fontes internas). Os dados são “reunidos de forma a explicitar a vida acadêmica do docente, das atividades de ensino, pesquisa e extensão universitária até as produções intelectuais e temáticas de pesquisa” (PORTAL..., 2019).

Na prática, “é um ambiente informacional digital que tem como objetivo dar visibilidade a dados públicos relacionados ao corpo docente, facilitando buscas das mais diversas dentro desse universo”, tais como: unidade em que o docente atua, tipo de vínculo que possui, área de atuação e detalhes de sua produção acadêmica e, de forma mais específica, dados de projetos financiados pela Fapesp ou pelo CNPq. Permite “encontrar especialistas especificando um tema, inserindo uma palavra-chave ou buscando por um programa de pós-graduação” (PORTAL..., 2019). As buscas[2] podem ser feitas por *Nome do docente* (2864), *Unidades/Reitoria* (35), *Unidades de lotação* (197), *Curso de graduação* (136), *Programas de pós-graduação* (153) e *Tópico/Assunto* (Figura 1):

[1] Disponível em: www.unesp.br/portaldocentes.

[2] Dados extraídos em setembro/2020.

Figura 1 – Página inicial do Portal Docentes Unesp



Fonte: Dados da pesquisa.

Quanto às buscas por *Tópico/Assunto*, isto é, à representação temática, podem ser realizadas, utilizando-se ou não as aspas (“). Com o uso das aspas, a recuperação é mais precisa, isto é, é recuperado exatamente o termo de busca digitado, sendo possível a visualização das informações sobre o(s) docente(s) que trabalha(m) nessa temática de pesquisa, ou seja, que usou(aram) a referida palavra-chave como palavra-chave de sua produção bibliográfica no Currículo Lattes. Para tanto, é calculado um percentual denominado *Aderência à busca*, quer dizer, o *nível de correspondência* ou *nível de consistência* entre a palavra-chave empregada na busca e a palavra-chave utilizada pelo docente, no preenchimento das palavras-chave do Currículo Lattes. Na prática, quanto mais usada a palavra-chave na busca pelo docente, em seu Currículo Lattes, maior o *nível de aderência* apresentado pelo Portal Docentes Unesp.

Um ponto interessante é a possibilidade de visualização de uma nuvem contendo as 15 palavras-chave mais utilizadas pelo pesquisador no Currículo Lattes, que “proporciona um panorama geral dos assuntos envolvidos nas pesquisas de cada docente e, no caso de interesse nas pesquisas desenvolvidas” (PORTAL..., 2019).

Nesse sentido, o Portal visa a auxiliar na localização de docentes ou pesquisadores da universidade, para “participar de missões ou atuar como parceiro de pesquisa em colaboração com instituições externas” (PORTAL..., 2019). No caso

de localização de informação incorreta ou desatualizada, no Portal, a orientação é para que o docente realize a atualização na base de dados de origem, já que os dados são extraídos automaticamente das fontes. Assim, a interação do docente com o Portal e o cuidado na qualidade dos dados possibilitam a concretização do objetivo de melhoria dos dados armazenados referentes à Unesp, nas diferentes bases, ensejando que o Portal "se torne uma ferramenta importante de uso para a comunidade interna e externa" (PORTAL..., 2019). Para o grupo de trabalho à frente do Portal Docentes Unesp, chama atenção "a diferença que faz o preenchimento cuidadoso dos formulários no momento de submeter um artigo científico, seja *na hora de listar palavras-chave do estudo* ou escrever de forma correta a filiação institucional" (PORTAL..., 2019, grifo nosso).

O Currículo Lattes, do CNPq, é a principal fonte de dados sobre os pesquisadores no Brasil. Foi idealizado nos anos 1980, por meio de um formulário-padrão para registro dos currículos, o qual permitisse "a avaliação curricular do pesquisador, a criação de uma base de dados que possibilitasse a seleção de consultores e especialistas, e a geração de estatísticas sobre a distribuição da pesquisa científica no Brasil" (HISTÓRIA..., 2020). Posteriormente, com o advento da Internet e as inovações tecnológicas, o Currículo Lattes, lançado oficialmente em 1999, tem sido aprimorado e "utilizado pelas principais universidades, institutos, centros de pesquisa e fundações de amparo à pesquisa dos estados como instrumento para a avaliação de pesquisadores, professores e alunos", no país e no exterior (HISTÓRIA ... 2020).

Possui 9 módulos: Dados gerais, Projetos, Produção bibliográfica, Produção técnica, Orientações, Produção cultural, Eventos, Bancas e Citações. O módulo "Produção bibliográfica", fonte de coleta da presente pesquisa, reúne informações sobre a produção bibliográfica (publicações) do pesquisador e subdivide-se nos seguintes tipos: a) *Artigos completos publicados em periódicos*: artigos científicos já publicados em revistas indexadas com ISSN; b) *Artigos aceitos para publicação;* artigos no prelo (in-press) que ainda não foram publicados em revistas indexadas com ISSN; c) *Livros e capítulos*: livros ou capítulos de livros produzidos, indexados com ISBN; d) *Texto em jornal ou revista*: qualquer publicação escrita que tenha sido publicada em meio

jornalístico, como roteiros, ensaios, matérias, reportagens, relatos, depoimentos, entrevistas, resumos, resenhas, crônicas, contos, poemas e afins; e) *Trabalhos publicados em anais de eventos*: textos publicados em anais de eventos, vinculados a um evento específico. Possui um vínculo com o item “Eventos”; f) *Apresentação de Trabalho*: apresentação de trabalho não vinculada a evento (aulas magnas, palestras, trabalhos acadêmicos etc.); g) *Partitura musical*: partituras escritas para canto, coral, orquestra etc.; h) *Tradução*: artigos, livros ou outras publicações traduzidas; i) *Prefácio*, *posfácio*: prefácio, posfácio, introdução ou apresentação de livros; j) *Outra produção bibliográfica*: qualquer outra produção bibliográfica que não se enquadre nas opções anteriores, inclusive artigos publicados em periódicos sem ISSN (CURRÍCULO ... 2020).

Embora o Módulo “Produção bibliográfica” seja talvez o mais importante, no Currículo Lattes, por possibilitar a visibilidade da produção acadêmica do docente, verifica-se a ausência de diretrizes/orientações aos autores, quanto à atribuição de palavras-chave de acordo com a temática do trabalho desenvolvido pelo pesquisador, isto é, ao correto preenchimento das palavras-chave de sua produção bibliográfica, na base de dados. No preenchimento das palavras-chave, é possível apenas adicionar palavras novas ou selecionar palavras já existentes na base de dados.

Na seleção dos docentes cujas palavras-chave dos Currículos Lattes seriam analisadas, utilizou-se a ferramenta de análise de dados SciVal[3], que se presta a realizar análises métricas da produção científica de uma instituição, país, região, autor ou grupos de autores, ou de periódicos adotando como fonte de dados a base de dados Scopus e o Science Direct, da Elsevier, cuja estrutura de metadados (perfis de autores, perfil institucional, lista de referências e assim por diante) permite a análise de mais de 30 conjuntos de métricas, de forma independente. Possui quatro módulos: a) *Overview* (Desempenho)*:* concentra-se no estado da arte da pesquisa, ou seja, em todo o panorama da situação atual de qualquer entidade/instituição selecionada, favorecendo o levantamento dos

[3] O acesso à ferramenta, assinada pela UNICAMP, foi realizado por meio do Portal do Sistema de Bibliotecas da UNICAMP (SBU), disponível em: http://www.sbu.unicamp.br/sbu/bases-de-dados/?op=2&ltr=Svier.

principais indicadores de performance; b) *Benchmarking:* possibilita a comparação entre instituições, grupos de pesquisa e pesquisadores; c) *Collaboration* (Colaboração)*:* permite a identificação de colaborações vigentes ou que já existiram, bem como colaborações internacionais em potencial, entre as entidades/instituições; d) *Trends* (Tendências): pressupõe a descoberta de áreas ou temas de pesquisa, em termos de proeminência mundial (SciVal, 2020).

No módulo “Overview”, selecionou-se a “Universidade Estadual Paulista Júlio de Mesquita Filho” como instituição a ser analisada, obtendo-se um resultado de 49.149 publicações; 488.365 citações recebidas pelas respectivas publicações, com uma média de 9,9 citações por publicação; 31.712 autores; 0,94 de impacto de citação ponderado por campo[4]; e Índice H5 (indexador h dos artigos publicados nos últimos cinco anos) de 86, entre 2010 e 2019, período máximo permitido de análise disponibilizado pela ferramenta[5] (Figura 2):

Figura 2 – Desempenho da Unesp na base de dados *Scopus*, entre 2010 e 2019

Fonte: Dados da pesquisa

No menu “Autores”, foi possível a visualização e exportação em Excel dos dados referentes aos primeiros 500 autores da Unesp que possuem produção

[4] Métrica utilizada pela Scopus e que se refere à proporção de citações recebidas em relação à média mundial esperada para o campo temático, tipo de publicação e ano de publicação. No caso da Unesp, 0,94 significa que as publicações foram citadas menos 6% do esperado. Por outro lado, “um impacto de citação ponderada no campo de mais de 1,00 indica que as publicações foram citadas mais do que seria esperado com base na média mundial de publicações semelhantes” (INDICADORES ..., 2020). Na Web of Science, essa métrica é denominada *impacto de citação normalizada*.

[5] Dados extraídos em setembro/2020.

bibliográfica na base de dados Scopus, ensejando o ranqueamento nas seguintes opções: *Produção acadêmica*; *Publicações mais recentes*; *Número de citações* e *Índice h* (Figura 3).

Figura 3 - Autores da Unesp com publicações na base de dados Scopus, entre 2010 e 2019.

Fonte: Dados da pesquisa

Tomando-se como base o ranqueamento do Índice h dos docentes, realizou-se a consulta dos autores no Portal Docentes Unesp, de modo a selecionar os 10 primeiros docentes que ainda possuem vínculo com a Universidade.

Após a seleção dos docentes, procedeu-se à busca dos respectivos Currículos Lattes no Portal do CNPq, a fim de se efetivar o levantamento das palavras-chave atribuídas pelos autores a suas produções bibliográficas. Para tanto, ao visualizar o Currículo Lattes do docente, selecionou-se a opção “Mostrar informações complementares” e, em seguida, percorreu-se o módulo “Produção bibliográfica”, de maneira a possibilitar a visualização e levantamento das palavras-chave atribuídas aos registros, organizadas no Excel. Fez-se o tratamento das palavras-chave, de sorte a verificar aspectos referentes à padronização e quais delas eram mais proeminentes, no universo de publicações e de docentes analisado.

Em um segundo momento, elegeu-se, da produção bibliográfica dos docentes, uma amostra de 280 artigos de periódicos, a fim de comparar as palavras-chave atribuídas pelos pesquisadores no Currículo Lattes e os

assuntos atribuídos no artigo original. Justifica-se a escolha desse tipo de produção bibliográfica por ser o meio principal de comunicação dos resultados de pesquisa, no meio científico. Para tanto, aplicou-se a metodologia da Avaliação da Indexação, na modalidade *Avaliação intrínseca quantitativa mediante a consistência*, que permite levantar medidas objetivas da qualidade da indexação por meio de fórmulas matemáticas. Nesta pesquisa, optou-se pela fórmula da consistência de Hooper (1965), adaptada por Gil Leiva (2008, p. 386) (Quadro 2):

Quadro 2 – Fórmula matemática para obtenção do índice de consistência entre as palavras-chave do Currículo Lattes e as do artigo original

$$C_i = \frac{T_{co}}{(A+B) - T_{co}}$$

Onde,

C_i = *Índice de consistência*
T_{co} = *Número de termos comuns nas duas indexações*
A = *Número de termos atribuídos na Indexação A* (Currículo Lattes)
B = *Número de termos atribuídos na Indexação B* (artigo original)

Fonte: Adaptado de Gil Leiva (2008, p. 386)

Na prática, há duas abordagens: *comparação rígida*, em que os assuntos atribuídos devem coincidir completamente, sendo considerado o valor de "1" para cada assunto atribuído nas duas indexações analisadas. Já a *comparação relaxada* (GIL LEIVA; RUBI; FUJITA, 2008, p. 238) ou *comparação flexível* (TARTAROTTI, 2014, p. 215) é empregada, quando há a coincidência total entre os assuntos atribuídos. Considera-se o valor de "1" ou "0,5", no caso de coincidência somente em uma parte do assunto. Tanto no índice rígido quanto no relaxado/flexível, quando não há coincidência entre os assuntos atribuídos, considera-se o valor de "0". Quando os valores são convertidos em porcentagem, a comparação dos termos varia de 0 a 100% de consistência (GIL LEIVA; RUBI; FUJITA, 2008, p. 238).

4 RESULTADOS

O levantamento das palavras-chave dos artigos de periódicos possibilitou duas frentes de análise: a) aplicabilidade da avaliação da indexação e b)

observação da padronização da apresentação das palavras-chave no Currículo Lattes dos docentes.

a) Avaliação da indexação

A *Avaliação intrínseca quantitativa mediante a interconsistência* permitiu a comparação entre as palavras-chave atribuídas pelos autores nos artigos de periódicos cadastrados no Currículo Lattes e os assuntos atribuídos aos respectivos artigos originais. Ao aplicarmos a fórmula da consistência de Hopper (1965), adaptada por Gil Leiva (2008, p. 386), como produtos "foram obtidos *ensaios* ou *índices de interconsistência*, verificando *inconsistências*, *discrepâncias* ou *diferenças* entre as duas indexações" (TARTAROTTI, 2019, p. 124) (Tabela 1):

Tabela 1 – Índices de interconsistência na indexação entre as palavras-chave do Currículo Lattes e o artigo original

Docente	Índice rígido	Índice flexível	Média
Docente 1	0%	0%	**0%**
Docente 2	0%	0%	**0%**
Docente 3	0%	0,81	**0,40%**
Docente 4	13,47%	14,25%	**13,86%**
Docente 5	100%	100%	**100%**
Docente 6	1,35%	2,04%	**1,69%**
Docente 7	9,43%	11,45%	**10,44%**
Docente 8	3,28%	4,05%	**3,66%**
Docente 9	1,90%	2,95%	**2,42%**
Docente 10	8,59%	9,58%	**9,08%**
Média por índice	**13,80%**	**14,51%**	-

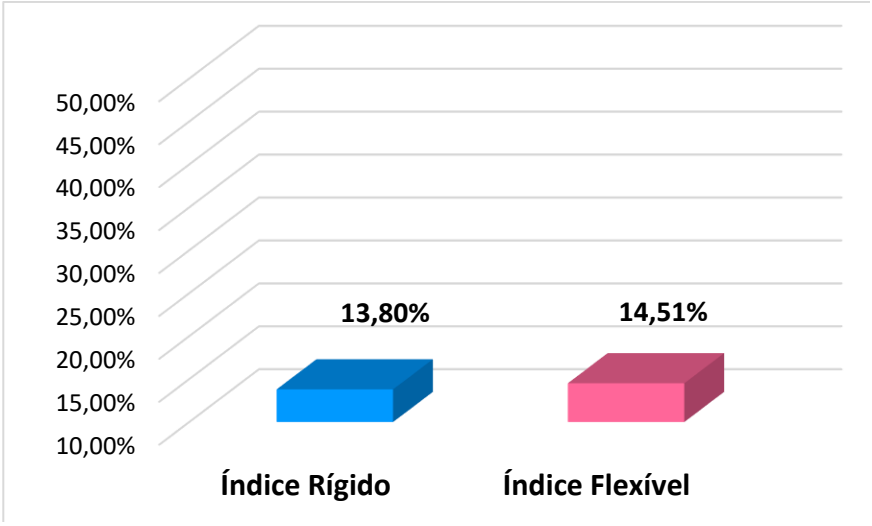
Fonte: Elaborada pelas autoras

Quanto à média dos índices de interconsistência por tipo de índice, observa-se que o maior nível foi obtido na comparação entre as palavras-chave do Currículo Lattes e os assuntos atribuídos no artigo original do *Docente 5*, com 100% de correspondência nos assuntos, entretanto, com apenas 2 artigos analisados. Em seguida, aparece o *Docente 4*, com 13,86%, o *Docente 7*, com 10,44%, o *Docente 10*, com 9,08%, o *Docente 8*, com 3,66%, o *Docente 9*, com 2,42%, o *Docente 6*, com 1,69% e o *Docente 3*, com apenas 0,40%. Tanto o *Docente 1* quanto o *Docente 2* aparecem com 0% de correspondência, isto é, a nenhum dos artigos de periódicos cadastrados em suas respectivas produções

científicas foram atribuídos novos termos ou replicadas as mesmas palavras-chave que constam nos artigos originais.

Em relação à média dos índices de interconsistência com o índice rígido e o índice flexível, não houve muita diferença entre ambos (Gráfico 1):

Gráfico 1 – Média dos índices de interconsistência na indexação entre as palavras-chave do Currículo Lattes e os artigos originais



Fonte: Elaborado pelas autoras

Um ponto a ser destacado é que em 47,5% do total de artigos originais dos docentes da amostra não houve atribuição de palavras-chave. Observou-se, entretanto, que palavras-chave foram atribuídas pelo autor para esses mesmos artigos cadastrados no Currículo Lattes. O docente 1, por exemplo, destacou-se com 100% dos artigos analisados com palavras-chave no Lattes, o que reforça a visão do autor como indexador.

No geral os resultados indicam baixos índices de interconsistência entre as palavras-chave atribuídas pelos docentes em artigos de periódicos cadastrados nos Currículos Lattes e os assuntos atribuídos nos artigos originais. Tal situação afeta diretamente a recuperação por assunto das produções científicas nos Currículos Lattes e no Portal Docentes da Unesp, tendo em vista que a aderência à busca é comprometida, assim como a apresentação fidedigna da nuvem de palavras-chave de cada docente, nesse ambiente informacional.

b) Padronização das palavras-chave

A análise e a compilação das palavras-chave atribuídas pelos docentes no Currículo Lattes permitiram, em um primeiro momento, a síntese da distribuição por docentes (Tabela 2):

Tabela 2 – Docentes e respectivos artigos de periódicos cadastrados no Currículo Lattes

Docente	Índice h	N. artigos	% artigos com palavras-chave
Docente 1	106	1227	4,56%
Docente 2	90	269	0%
Docente 3	95	953	6,82%
Docente 4	53	465	60,21%
Docente 5	48	227	0,88%
Docente 6	45	389	98,71%
Docente 7	42	262	68,32%
Docente 8	42	217	22,58%
Docente 9	41	196	95,92%
Docente 10	40	344	47,38%
Média	**60,2**	**455**	**40,54%**

Fonte: Elaborada pelas autoras

Observa-se que a média dos 10 docentes com maiores índices h da Unesp corresponde a 60,2, sendo o maior índice o do *Docente 1* (106) e o menor índice, o do *Docente 10* (40). Quanto ao número de artigos, percebe-se que o docente com a maior quantidade de artigos cadastrados no Currículo Lattes é o *Docente 1*, com 1227 artigos, enquanto o *Docente 9* se encontra com a menor quantidade (196), com uma média de 455 artigos por docente. Salienta-se que os docentes analisados são das áreas do conhecimento de Exatas e Biológicas, fato que explica o expressivo número de artigos publicados.

Em relação à porcentagem de artigos com palavras-chave cadastradas no Currículo Lattes, nota-se que o *Docente 6* possui o maior índice, com 98,71% de preenchimento, ao passo que o *Docente 2* aparece com 0%, isto é, não houve atribuição de palavras-chave nesse tipo de produção bibliográfica.

Dentre as palavras-chave cadastradas no Currículo Lattes, realizou-se um comparativo entre as 10 que mais se repetem e a nuvem de palavras obtida na busca pelo docente no Portal Docentes Unesp (Tabela 3):

Tabela 3 – Palavras-chave mais proeminentes nos artigos de periódicos do Currículo Lattes dos docentes

Docente	Palavra-chave	N.
Docente 1	Acoplamentos anômalos	11
	Boson de Higgs	6
	Boson Higgs	4
	Higgs Boson	4
	Colisoes Eletron-Foton	4
	Supersimetria	3
	Espalhamento de Gravitons	3
	Colisoes E^+ E^-	3
	Colisoes Eletron-Foton	3
	Correcoes Radiativas	3
	...	...
Docente 2	-	-
Docente 3	SUSY search in pp collisions at LHC with CMS	5
	Fenomenologia do Efeito Hanbury-Brown Twiss	4
	Modelo Covariante Geral Para A Interferometria	3
	Correlacoes back-to-back entre fermion-antifermion	2
	Correlacoes Back-to-Back entre boson-antiboson	2
	Correla ├º ├Áes Hadr ├┤ nicas Comprimidas-modos de busca	2
		2
	Medidas de pT e pseudo-rapidez no detector CMS	2
	Se	2
	Search for scalar leptoquarks at LHC with CMS	1
	1-D HBT in MB and HM events in pp collisions-CMS	...
	...	
Docente 4	Glasses	48
	bacterial cellulose	16
	Glass	13
	luiminescence	11
	Crystallization	8
	Infrared	8
	photoluminescence	8
	energy transfer	6
	glass-ceramic	5
	energy-transfer	4
	...	...
Docente 5	Noise Control	1
	robust control	1
	Optimum control	1
	Structural Health Monitoring – SHM	1
	Analysis of Variance – ANOVA	1
	Impedância Eletromecânica	1
Docente 6	Anura	187
	Amphibia	82
	Taxonomia	59
	Atlantic Forest	41
	Conservation	41
	New Species	35
	Distribution	28
	Phylogeny	22
	Vocalizations	20
	Biodiversity	18
	...	...

Docente 7	Eletroanalitica	13
	dye contaminants	12
	cathodic stripping voltammetric	11
	electrochemical sensor	11
	photoelectrochemistry	11
	Modified Electrode	9
	Azo Dye	8
	electroanalytical method	8
	Reactive Dyes	8
	Dyes Determination	7
	...	...
Docente 8	Atlantic Forest	23
	Conservacao	8
	Dispersao de Sementes	6
	Seed Dispersal	6
	Seed Predation	6
	Diet	5
	Ecologia de Primatas	5
	Predacao de Sementes	5
	Frugivoria	4
	Frugivory	4
Docente 9	Impedance Spectroscopy	44
	varistor	40
	SnO2	39
	Capacitance Spectroscopy	18
	Biosensor	15
	Electrochemical Impedance Spectroscopy	15
	CCTO	12
	Electrochromism	12
	TiO2	12
	ZnO	12
	...	...
Docente 10	Rubiaceae	14
	Iridoids	4
	Pterogyne nitens	7
	piperaceae	5
	Piper tuberculatum	5
	antifungal activity	5
	Alibertia macrophylla	4
	Alzheimer	4
	antifungal compounds	4
	antioxidant	4
	...	...

Fonte: Elaborada pelas autoras

O levantamento das nuvens de palavras-chave dos docentes disponíveis no Portal Docentes Unesp possibilitou a visualização das palavras-chave que mais se repetem, na produção científica dos docentes, cadastradas nos artigos de periódicos do Currículo Lattes, conforme apresentado a seguir. As palavras-chave com as maiores fontes correspondem às palavras-chave que mais apareceram na produção científica dos pesquisadores, entre artigos completos publicados em periódicos, livros publicados/organizados ou edições, capítulos

de livros publicados, livros, textos em jornais de notícias/revistas, trabalhos completos publicados em anais de congressos, resumos expandidos publicados em anais de congressos, resumos publicados em anais de congressos, resumos publicados em anais de congresso (artigos) e artigos aceitos para publicação.

Na nuvem de palavras-chave do *Docente 1* (Figura 4), verifica-se a ausência de padronização nos termos “Boson de Higgs”, “Higgs Boson” e “Boson Higgs”, os quais figuram entre os 10 mais utilizados nos artigos de periódicos. Outros exemplos observados são: “Acoplamento anomalo” e “Acoplamentos anomalos”; “Colisoes E\Gamma e \Gamma\Gamma” e “Colisoes E-Gamma”; “Lagrangeanas Efetivas” e “Lagrangeans Efetivas”; “Particulas de Spin 2” e “Particulas de Spin 3/2”.

Figura 4 – Nuvem de palavras-chave – Docente 1

Fonte: Dados de pesquisa

Na nuvem de palavras do *Docente 2* (Figura 5), a fonte das palavras-chave é oriunda de outras produções bibliográficas, já que nenhum dos artigos periódicos cadastrados no Currículo Lattes possui as palavras-chave cadastradas.

Figura 5 – Nuvem de palavras-chave – Docente 2

Fonte: Dados de pesquisa

Na nuvem de palavras do *Docente 3* (Figura 5), observa-se uma maior quantidade de diferentes palavras-chave, com um menor índice de inconsistência nos termos, como, por exemplo, “Correla ├º ├úo BBC entre pares de meson phi” e “Correla ├º ├Áes Comprimidas entre pares de mesons phi”.

Figura 5 – Nuvem de palavras-chave – Docente 3

Fonte: Dados de pesquisa

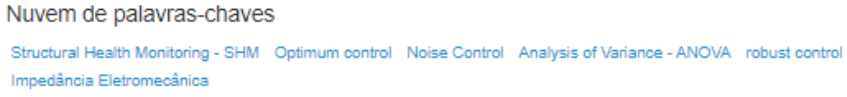
Na nuvem de palavras do *Docente 4* (Figura 6), verifica-se a ausência de padronização nos termos “Glasses” e “Glass”, os quais figuram entre os 10 mais empregados nos artigos de periódicos. Outros exemplos observados são: “bacterial”, “bacterial cellulose”, “cellulose” e “celulose bacteriana”; “energy transfer” e “energy-transfer”; “Eu3+” e “Eu3+ And Eu2+”; “Europio” e “europium”; “fluoroindate” e “fluoroindate glasses”; “glass ceramic” e “glass-ceramics”; “laser” e “lasers”; “luminescence” e “Luminescencia”; “mechanism” e “mechanisms”; “Nanocomposite” e “Nanocomposites”; “Optical properties” e “optical-properties”; “Organic-Inorganic Hybrid (OIHs)” e “organic-inorganic hybrids”; “sol-“, “sol gel”, “sol gel route”, “sol-gel”, “sol-gel method” e “Sol-Gel Synthesis”; “Tellurite”, “tellurite glass” e “tellurite glasses”; “hybrids waveguides” e “waveguides”.

Figura 6 – Nuvem de palavras-chave – Docente 4

Fonte: Dados de pesquisa

Já na nuvem de palavras do *Docente 5* (Figura 7), a fonte das palavras-chave são outras produções bibliográficas, tendo em vista que apenas 2 artigos de periódicos continham as palavras-chave cadastradas no Currículo Lattes.

Figura 7 – Nuvem de palavras-chave – Docente 5



Fonte: Dados de pesquisa

Na nuvem de palavras do *Docente 6* (Figura 8), tem-se a ausência de padronização nos termos “Conservation” e “Conservacao”, que figuram entre os 10 mais usados nos artigos de periódicos. Outros exemplos são: “Amphibia” e “Anfibios”; “Anura”, “Anurans” e “Anuros”; “Behavior” e “Comportamento”; “Biodiversidade” e “Biodiversity”; “Brasil” e “Brazil”; “Breeding” e “Breeding activity”; “Cariotipo” e “Cariotipos”; “Comunidade”, “Comunidade de Anuros” e “Comunidade Anura”; “Declines” e “Declinios”; “Distribuicao”e “Distribution”; “Diversidade” e “Diversity”; “Evolução” e “Evolution”; “Floresta Atlantica”, “Atlantic Forest” e “Mata Atlantica”; “Ecologia” e “Ecology”; “Girino” e “Girinos”; “Hyla” e “Hylidae”; “Karyotype” e “Karyotypes”; “New Species” e “Novas Especies”; “Reprodução”, “Reproducao Animal”, “Reproduction”, “Reproductive Biology” e “Reproductive Mode”; “SistemÃ¡tica Molecular” e “Sistematica”; “Vocalização”, “Vocalizaciones”, “Vocalizations” e “Vocalizatio”.

Figura 8 – Nuvem de palavras-chave – Docente 6

Fonte: Dados de pesquisa

Nas palavras-chave atribuídas nos artigos cadastrados no Currículo Lattes pelo *Docente 7*, verifica-se a ausência de padronização nos termos: “Alfa, Beta-Insaturated Carbonyls” e “Alfa-Beta-Insatured Carbonyls”; “Azo Dye” e “Azo Dyes”; “disperse azo dye” e “disperse dyes analysis”; “biological treatement” e “biological treatment for dyes”; “cathodic stripping voltammetric”, “Cathodic Stripping Voltammetry” e “Cathodic Stripping Voltammmetry dye”; “marker dye”, “marker analysis” e “Dyes Determination”; “Electroanalysis”, “electroanalytical method”, “Eletroanalise” e “Eletroanalitica”; “Eletroquimica” e “Eletroquimica Organica”; “FOOD DYE”, “food dye analysis fuel” e “fuel analysis”; “hair dye” e “hair dyes”; “high performance liquid chromatography” e “HPLC- ED”; “Pharmaceutical Drug” e “Pharmaceutical Drugs”; “Polyamino Acid” e “Poly-Aminoacids”; “reduction of CO2” e “co2 reduction”; “Voltammetric Analysis” e “Voltammetric Determination”. Nuvem de palavras-chave (Figura 9):

Figura 9 – Nuvem de palavras-chave – Docente 7

Fonte: Dados de pesquisa

Nas palavras-chave atribuídas nos artigos cadastrados no Currículo Lattes pelo *Docente 8*, nota-se a ausência de padronização nos termos “Conservation” e “Conservacao”, mesma ocorrência para as do *Docente 6*, que figuram entre as 10 palavras-chave mais utilizadas nos artigos de periódicos. Na nuvem de palavras-chave (Figura 10) também aparecem “Atlantic forest” e “Mata Atlântica”; “Seed dispersal” e “Dispersão de sementes” e “Seed predation” e “Predação de sementes”, estes também correspondentes a “Ecologia da Dispersao de Sementes”. Outros exemplos são: “Frugivores”, “Frugivoria” e “Frugivory”; “Migracao” e “Migration”; “Ecologia de Primatas” e “Primates”; “Migracao” e “Migration”; “Psittacidae” e “Psittacideos”; “Rodentia” e “Rodents”, “Aves” e “Birds”.

Figura 10 – Nuvem de palavras-chave – Docente 8

Fonte: Dados de pesquisa

Nas palavras-chave atribuídas nos artigos cadastrados no Currículo Lattes pelo *Docente 9*, observa-se a ausência de padronização nos termos “Impedance spectroscopy” e “Electrochemical impedance spectroscopy”, que figuram entre as 10 palavras-chave mais usadas nos artigos de periódicos. Outros exemplos são: “Bioelectroanalys” e “Bioelectroanalysis”; “Biosensor”, “Biosensores” e “Biosensors”; “biotechnology” e “Biotecnologia”; “Capacitance Spectroscopy” e “Capacitance Spectroscpy”; “Dielectric properties” e “dielectric property DSC (Dye Sensitized Solar Cell)”; “Dye-Sensitized Solar Cell

impedance” e “Impedance Spectroscopy”; “Insertion” e “Insertion electrodes”; “kinetic” e “Kinetics”; “lectin” e “Lectinas”; “Nanostructred electrodes” e “Nanostructured electrodes”; “Nonohimic devices”, “nonohmic devices” e “nonohmic properties”; “polycrystalline semiconductor” e “polycrystalline semiconductors”; “QCM” e “Quartz Crystal Microbalance”; “varistor” e “Varistores”; “Zinc oxide” e “ZnO”. Nuvem de palavras-chave (Figura 11):

Figura 11 – Nuvem de palavras-chave – Docente 9

Fonte: Dados de pesquisa

Nas palavras-chave atribuídas nos artigos cadastrados no Currículo Lattes pelo *Docente 10*, verifica-se a ausência de padronização nos termos “antioxidant” e “antioxidante”, os quais figuram entre os 10 mais utilizados nos artigos de periódicos e que também correspondem a outros termos: “Antioxidant”, “Antioxidant activity”, “Antioxidante” e “antioxidantes”. Outros exemplos são: “acetilcolinesterase”, “acetylcholinesterase”, “acetylcholinesterase (AChE) inhibitors” e “Acetylcholinestesae inhibitors”; “Antibacterial Activity” e “Antibacterial”; “Antifungal” e “antifungal activity”; “BIODIVERSIDADE” e “Biodiversity”; “Candida”, “CANDIDA SP” e “Candida ssp”; “Cassia” e “Cassia spectabilis”; “lignanas” e “Lignans”; “Penicillium” e “Penicillium sp”; “Piper aduncum” e “Piper aduncum L.”; “Piperacae” e “piperaceae”. Nuvem de palavras-chave (Figura 10):

Figura 10 – Nuvem de palavras-chave – Docente 10

Fonte: Dados de pesquisa

Observa-se que a ausência de padronização na atribuição das palavras-chave na produção bibliográfica dos docentes no Currículo Lattes leva à sua separação na nuvem, prejudicando a adequada recuperação por temática de pesquisa, visto que, se reunidos, os termos apareceriam de forma mais proeminente na nuvem, possibilitando a visualização de outras palavras-chave não contempladas.

5 CONCLUSÕES

Os resultados de baixos índices de interconsistência entre as palavras-chave, no Currículo Lattes dos docentes, em artigos de periódicos, e os assuntos atribuídos nos artigos originais indicam ausência de uma política de indexação formalizada, a qual forneça diretrizes aos docentes/pesquisadores na atribuição padronizada das palavras-chave em suas produções bibliográficas, no Currículo Lattes, com base em uma linguagem de indexação controlada/padronizada.

Quanto aos aspectos sintáticos e semânticos das palavras-chave, verifica-se, de forma geral, que as maiores inconsistências estão ligadas à escolha em português e/ou inglês, à adoção de singular ou plural, erros de grafia, palavras-chave em caixa alta ou em caixa baixa, em sua totalidade etc. Também foi possível verificar, na extração outros dados, que não correspondem a metadados de assuntos, relacionados a meio de divulgação, ISBN/ISSN, Homepage do trabalho e Série.

Esses resultados, os quais influenciam a qualidade de funcionalidade para outras aplicações, nos levam a considerar dois aspectos prováveis a respeito da atribuição de palavras-chave por autores: da alta especificidade de novos termos e da alta incompatibilidade de termos, por problemas ortográficos e semânticos. O primeiro aspecto diz respeito a áreas como, por exemplo, a biologia molecular, com rápida atualização a partir de novas pesquisas que utilizam novos métodos e técnicas para a descoberta de novos resultados (HAN; HARRINGTON; BLACK; KUDEKI, 2016). O segundo aspecto está associado à falta de padronização e controle de vocabulário, que gera variações terminológicas no uso de diferentes grafias para um mesmo termo, como, por exemplo, o termo composto “Macaco prego” que, com a colocação de um hífen, “Macaco-prego”

(FUJITA; TOLARE, 2019), gera um outro termo, ou mesmo quando se atribui um termo no singular, “Mulher”, ou, no plural, “Mulheres”.

Se, por um lado, existem problemas com palavras-chave, por outro, com certeza, deverá haver problemas com os vocabulários controlados que não conseguem acompanhar a evolução de novos termos ou termos específicos. De qualquer modo, é necessário avançar na concepção de indexadores não proficientes e elaborar proposta de política de indexação para padronização de palavras-chave atribuídas por autores e pesquisadores, na submissão da produção científica em diferentes sistemas de informação que realizam a gestão e divulgação científica.

A solução de problemas de representação e recuperação deve ser estudada, após avaliações de diferentes perspectivas, conforme características de cada sistema de armazenagem e recuperação da informação. Contudo, será necessário discutir a elaboração de uma política de organização e representação da informação a ser seguida, que possa ser continuamente avaliada e atualizada. No caso do Portal Docentes Unesp, recomenda-se dar ciência aos docentes dos resultados desta pesquisa, para que possam corrigir suas palavras-chave de artigos, à luz de orientações de padronização e consistência.

REFERÊNCIAS

BALATSOUKAS, P.; ROUSIDIS, D.; GAROUFALLOU, E. A method for examining metadata quality in open research datasets using the OAI-PMH and SQL queries: the case of the Dublin Core ‘Subject’ element and suggestions for user-centred metadata annotation design**. International Journal of Metadata, Semantics and Ontologies**, [s.l.], v. 13, n. 1, p. 1-8, 2018. Disponível em: doi:10.1504/IJMSO.2018.096444. Acesso em: 28 set. 2020.

BOGERS, T.; PETRAS, V. Tagging vs. Controlled vocabulary: which is more helpful for book search? *In:* iCONFERENCE 2015: create, collaborate, celebrate, 2015, Newport Beach. **Proceeedings** [...]. Newport Beach, CA, 2015. p. 1-15. Disponível em: https://www.ideals.illinois.edu/bitstream/handle/2142/73673/65_ready.pdf?sequence=2. Acesso em: 12 set. 2020.

CHU, H. **Information representation and retrieval in the digital age**. 2. ed. Medford, NJ: Information Today, 2010. 306 p. (ASIST Monographs Series).

CURRÍCULO Lattes. **Portal do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)**, 2020. Disponível em: http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar. Acesso em: 28 set. 2020.

FOSKETT, A. C. **A abordagem temática da informação**. Trad. de Antonio Agenor Briquet de Lemos. São Paulo: Polígono; Brasília: Universidade de Brasília, 1973. 437 p.

FROTA, M. G. C. Memória e registro das violações aos direitos da criança nos documentos da corte interamericana de direitos humanos. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [s.l.], v. 7, n. 1, 2014. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119497. Acesso em: 30 set. 2020.

FUJITA, M. S. L. A representação documentária de artigos científicos em educação especial: orientação aos autores para determinação de palavras chaves. **Revista Brasileira de Educação Especial**, Marília, v. 10, n. 3, p. 257-272, set.-dez. 2004. Disponível em: https://www.researchgate.net/publication/277102627_A_representacao_documentaria_de_artigos_cientificos_em_educacao_especial_orientacao_aos_autores_para_determinacao_de_palavras_chaves. Acesso em: 15 set. 2020.

FUJITA, M. S. L.; AGUSTÍN-LACRUZ; M.-D.-C.; TERRA, A. L. Journals' guidelines about title, abstract and keywords: an overview of Information Science and Communication Science areas. **European Science Editing**, [s.l.], v. 44, p. 76–79, nov. 2018a. Disponível em: https://europeanscienceediting.eu/articles/journals-guidelines-about-title-abstract-and-keywords-an-overview-of-information-science-and-communication-science-areas/. Acesso em: 28 set. 2020.

FUJITA, M. S. L.; AGUSTÍN-LACRUZ, M.-D.-C; TERRA, A. L. Knowledge organization in editorial policies for titles, abstracts and keywords in JCR-indexed journals: an exploratory study in the areas of Information and Communication Sciences. *In*: RIBEIRO, F.; CERVEIRA, M. E. *In:* RIBEIRO, F.; CERVEIRA, M. E. Challenges and Opportunities for Knowledge Organization in the Digital Age: INTERNATIONAL ISKO CONFERENCE, 15., Porto, Portugal. **Proceedings** [...] Baden-Baden: Ergon, 2018b. p. 321-330. (Advances in Knowledge Organization, v. 16).

FUJITA, M. S. L.; TOLARE, J. Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros. **Informação & Informação**, Londrina, v. 24, p. 93-125, 2019. Disponível em: doi: 10.5433/1981-8920.2019v24n2p93. Acesso em: 15 set. 2020.

GARCIA, D. C. F.; GATTAZ, C. C.; GATTAZ, N. C. A relevância do título, do resumo e de palavras-chave para a escrita de artigos científicos. **Revista de Administração Contemporânea**, [s.l.], v. 23, n. 3, maio/junho, 2019.

Disponível em: http://www.scielo.br/pdf/rac/v23n3/1982-7849-rac-2019190178.pdf. Acesso em: 30 jun. 2020.

GIL LEIVA, I. **Manual de indización***:* teoría y práctica. Gijón: Trea, 2008.

GIL-LEIVA, I.; ALONSO-ARROYO, A. Keywords given by authors of scientific articles in database descriptors. **Journal of the American Society for Information Science and Technology**, [s.l.], v. 58, n. 8, p. 1175–1187, 2007. Disponível em: doi: 10.1002/asi.20595. Acesso em: 12 set. 2020.

GIL LEIVA, I.; RUBI, M. P.; FUJITA, M. S. L. Consistência na indexação em bibliotecas universitárias brasileiras. **Transinformação**, Campinas, v. 20, n. 3, p. 233-253, set./dez. 2008.

GOLUB, K.; TYRKKÖ, J.; HANSSON, J.; AHLSTRÖM, I. Subject indexing in humanities: a comparison between a local university repository and an international bibliographic service. **Journal of Documentation**, [s.l.], v. 76, n. 6, p. 1193-1214, 2020. Disponível em: doi: 10.1108/JD-12-2019-0231. Acesso em: 11 set. 2020.

GONÇALVES, A. L. Uso de resumos e palavras-chave em Ciências Sociais: uma avaliação. **Encontros Bibli:** revista eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 13, n. 26, p. 78-93, out. 2008. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2008v13n26p78. Acesso em: 28 set. 2020.

GROSS, T.; TAYLOR, A. G. What have we got to lose? The effect of controlled vocabulary on keyword searching results. **College & Research Libraries**, [s.l.], v. 66, n. 3, p. 212-230, may 2005. Disponível em: doi: https://doi.org/10.5860/crl.66.3.212. Acesso em: 12 set. 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. Disponível em: doi: 10.1080/01639374.2014.917447. Acesso em: 01 out. 2020.

GULL, C. D. Information science and technology: from coordinate indexing to the global brain. **Journal of the American Society for Information Science**, [s.l.], v. 38, n. 5, p. 338-66, 1987.

HAN, M.-J. K.; HARRINGTON, P.; BLACK, A.; KUDEKI, D. Aligning author-supplied keywords for ETDS with domain-specific controlled vocabularies. *In:* CLASSIFICATION AND INDEXING SATELLITE CONFERENCE, 2016, Ohio, Columbus. **Anais** […]. Ohio, Columbus: State Library of Ohio, 2016, p. 1-10. Disponível em: http://hdl.handle.net/2142/97879. Acesso em: 10 set. 2020.

HANRATH, S.; RADIO, E. User search terms and controlled subject vocabularies in an institutional repository. **Library Hi Tech**, [s.l.], v. 35, n. 3, p.

360-367, 2017. Disponível em: doi: 10.1108/LHT-11-2016-0133. Acesso em: 11 set. 2020.

HIDER, P. The retrieval power added by subject indexing to bibliographic databases. *In:* RIBEIRO, F.; CERVEIRA, M. E. Challenges and Opportunities for Knowledge Organization in the Digital Age: INTERNATIONAL ISKO CONFERENCE, 15., Porto, Portugal. **Proceedings** [...] Baden-Baden: Ergon, 2018. p. 426-431. (Advances in Knowledge Organization, v. 16).

HISTÓRIA do surgimento da Plataforma Lattes. **Portal CNPq**. Disponível em: https://bityli.com/krDV2. Acesso em: 29 set. 2020.

HORN, M. E. “Garbage” in, “refuse and refuse disposal” out: making the most of the subject authority file in the OPAC. **Library Resources & Technical Services**, [s.l.], v. 46, n. 3, p. 92-102, jul. 2002. Disponível em: doi: 10.1007/s12109-018-9590-3. Acesso em: 02 out. 2020.

INDICADORES e métricas. **Águia**: Agência USP de Gestão da Informação Acadêmica da Universidade de São Paulo, 2020. Disponível em: https://bityli.com/7I0cr. Acesso em: 28 set. 2020.

INTERNATIONAL Federation of Library Associations and Institutions (IFLA). **IFLA Library Reference Model**: um modelo conceitual para a informação bibliográfica: definição de um modelo de referência conceitual para fornecer uma estrutura para a análise de metadados não administrativos relacionados aos recursos das bibliotecas. Tradução de Isabel Cristina Ayres da Silva Maringelli, José Fernando Modesto da Silva, Liliana Giusti Serra, Luiza Wainer, Marcelo Votto Texeira, Raildo de Sousa Machado e Zaira Regina Zafalon. Título original: RIVA, P.; LE BOEUF, P.; ŽUMER, M. IFLA Library Reference Model: a conceptual model for bibliographic information: definition of a conceptual reference model to provide a framework for the analysis of non-administrative metadata relating to library resources. Revised after world-wide review; endorsed by the IFLA Professional Committee. 2017. 101 p. Disponível em: https://www.ifla.org/publications/node/11412. Acesso em: 29 set. 2020.

JENNIFER, P.; MUTHUKUMARAVEL, J. P. Indexing on IR system by sing stemming and stopwords. **International Journal of Recent Technology and Engineering**, [s.l.], v. 8, n. 1S2, p. 281-283, mai. 2019. Disponível em: https://www.ijrte.org/wp-content/uploads/papers/v8i1S2/A00650581S219.pdf. Acesso em: 28 set. 2020.

JOORABCHI A., MAHDI, A. E. Automatic subject metadata generation for scientific documents using wikipedia and genetic algorithms. *In*: TEN TEIJE A.; VOLKER, J.; HANDSCHUH, S.; STUCKENSCHMIDT, H.; D´ACQUIN, M.; NIKOLOV, A.; AUSSENAC-GILLES, N.; HERNANDEZ, N. (ed.). INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT (EKAW), 2012, Berlin, **Proceeedings** [...]. Lecture Notes in Computer Science. Berlin: Springer, v. 7603, 2012. Disponível em: https://doi.org/10.1007/978-3-642-33876-2_6. Acesso em: 28 set. 2020.

KIPP, M. User, author and professional indexing in context: an exploration of tagging practices on CiteULike**. Canadian Journal of Information and Library Science**, [s.l.], v. 35, n. 1, p. 1-41, 2009. Disponível em: https://www.researchgate.net/publication/46132798_User_Author_and_Professional_Indexing_in_Context_An_Exploration_of_Tagging_Practices_on_CiteULike. Acesso em: 15 set. 2020.

KIYOTA, Y.; TAMURA, N.; SAKAI, S.; NAKAGAWA, H.; MASUDA, H. Automated subject induction from query keywords through wikipedia categories and subject headings. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2008. **Proceedings** [...]. Marrakech: European Language Resources Association (ELRA), 2008. Disponível em: http://www.lrec-conf.org/proceedings/lrec2008/pdf/882_paper.pdf. Acesso em: 15 set. 2020.

KUN, L.; KIPP, M. E. J. Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: an experimental study on medical collections. **Journal of the Association for Information Science and Technology**, [s.l.], v. 65, n. 3, p. 483–500, 2014. Disponível em: doi: 10.1002/asi.22985. Acesso em: 20 set. 2020.

LANCASTER, F. W**. Indexação e resumos:** teoria e prática. 2. ed. rev. e atual. Trad. de Antonio Agenor de Briquet de Lemos. Brasília: Briquet de Lemos, 2004. 452 p.

LI, M. Classifying and ranking topic terms based on a novel approach: role differentiation of author keywords. **Scientometrics**, v. 116, p. 77–100, 2018. Disponível em: doi: 10.1007/s11192-018-2741-7. Acesso em: 21 set. 2020.

LU, W.; LI, X.; ZHIFENG, L.; CHENG, Q. How do author-selected keywords function semantically in scientific manuscripts? **Knowledge Organization**, [s.l.], v. 46, n. 6, p. 403-18, 2019. Disponível em: doi: 10.5771/0943-7444-2019-6-402. Acesso em: 10 set. 2020.

MAURER, M. B.; SHAKERI, S. Disciplinary differences: LCSH and keyword assignment for ETDS from different disciplines. **Cataloging & Classification Quarterly**, [s.l.]. v. 54, n. 4, p. 213-43, 2013. Disponível em: doi: 10.1080/01639374.2016.1141133. Acesso em: 20 set. 2020.

MÓDULO Produção bibliográfica. **Portal CNPq**. Disponível em: https://bityli.com/YPVIk. Acesso em: 29 set. 2020.

MOULAISON, H. L. Social tagging in the web 2.0 environment: author vs. user tagging. **Journal of Library Metadata**, [s.l.], v. 8, n. 2, p. 101-111, 2008. Disponível em: doi: 10.1080/10911360802087325. Acesso em: 15 set. 2020.

MUNAN, L. Classifying and ranking topic terms based on a novel approach: role differentiation of author keywords. **Scientometrics**, [s.l.], v. 116**,** p. 77–100,

2018. Disponível em: doi: 10.1007/s11192-018-2741-7. Acesso em: 01 out. 2020.

NÉVÉOL, A.; DOĞAN, R. I.; ZHIYONG, L. Author keywords in biomedical journal articles. *In*: AMIA ANNUAL SYMPOSIUM, 2010. **Proceedings** [...]. 2010, p. 537-541. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041277/. Acesso em: 28 set. 2020.

NEWMAN, D.; HAGEDORN, K.; SMYTH, C. C. P. Subject metadata enrichment using statistical topic models. *In*: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL), 2007, Vancouver. **Proceedings** [...]. Vancouver, BC, Canada, 2007. Disponível em: doi: 10.1145/1255175.1255248. Acesso em: 15 set. 2020.

PESET, F.; GARZÓN-FARINÓS, F.; GONZÁLEZ, L. M.; GARCÍA-MASSÓ, X.; FERRER-SAPENA, A.; TOCA-HERRERA, J. L.; SÁNCHEZ-PÉREZ E.A. Survival analysis of author keywords: an application to the library and information sciences area. **Journal of the Association for Information Science and Technology**, [s.l.], v. 71, n. 4, p. 462–473, 2020. Disponível em: doi: 10.1002/asi.24248. Acesso em: 20 set. 2020.

PORTAL docentes Unesp é lançado com dados de 3.000 professores. **Portal Unesp,** 27 jun. 2019. Disponível em: https://www2.unesp.br/portal#!/noticia/34769/portal-docentes-unesp-e-lancado-com-dados-de-3000-professores. Acesso em: 25 set. 2020.

ROWLEY, J. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. **Journal of Information Science**, [s.l.], v. 20, n. 2, p. 108-118, 1994. Disponível em: doi: 10.1177/016555159402000204. Acesso em: 18 set. 2020.

SANTOS, J. C. F. dos; CERVANTES, B. M. N. Controle de vocabulário em periódicos científicos eletrônicos. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais** [...]. João Pessoa: UFPB, 2015. p. 1-21. Disponível em: http://www.ufpb.br/evento/index.php/enancib2015/enancib2015/paper/viewFile/3068/997. Acesso em: 18 set. 2020.

SANTOS, R. F. dos; CORRÊA, R. F. Organização da informação em repositórios digitais: uma abordagem sobre a política de indexação da base de dados referencial de artigos de periódicos em Ciência da Informação (BRAPCI). *In:* GUIMARÃES, J. A. C. (org.). **Memória, tecnologia e cultura na organização do conhecimento**. Recife: Ed. UFPE, 2017. 409 f. (Série: Estudos Avançados em Organização do Conhecimento, v. 4). p. 249-261. Disponível em: http://hdl.handle.net/20.500.11959/brapci/122077. Acesso em: 28 set. 2020.

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. Disponível em: doi: 10.1080/01639374.2012.703164. Acesso em: 15 set. 2020.

SCIVAL. Disponível em: https://www.elsevier.com/solutions/scival. Acesso em: 10 out. 2020.

SMIRAGLIA, R. P. Keywords, indexing, text analysis: an editorial. **Knowledge Organization**, v. 40, n. 3, p. 155-9, 2013.

TARTAROTTI, R. C. D. **Atuação bibliotecária no tratamento temático da informação em unidades informacionais**: um estudo comparativo qualitativo-quantitativo. 2014. 277 f. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) – Universidade Federal de São Carlos, São Carlos, 2014. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/1140/6320.pdf?sequence=1&isAllowed=y. Acesso em: 12 out. 2020.

TARTAROTTI, R. C. D. **Avaliação do processo de indexação de assuntos em repositórios institucionais pela abordagem da recuperação da informação**. 2019. 370 f. Tese (Doutorado em Ciência da Informação - Faculdade de Filosofia e Ciências) – Universidade Estadual Paulista, Marília, 2019. Disponível em: https://repositorio.unesp.br/handle/11449/191064. Acesso em: 11 set. 2020.

TOEPFER, M.; SEIFERT, C. Descriptor-invariant fusion architectures for automatic subject indexing. *In*: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES, 2017, Canadá. **Proceedings** [...]. Toronto, ON, Canadá, 2017. p. 1-10. Disponível em: doi: 10.1109/JCDL.2017.7991557. Acesso em: 18 set. 2020.

VANYUSHKIN, A.; GRASCHENKO, L. Analysis of text collections for the purposes**. Journal of Information and Organizational Sciences**, [s.l.], v. 44, n.1, p. 171-184, 2020. Disponível em: doi: 10.31341/jios.44.1.8. Acesso em: 12 set. 2020.

WILLIS, C.; LEE, R. M. A random walk on an ontology: using thesaurus structure for automatic subject indexing. **Journal of the American Society for Information Science and Technology**, [s.l.], v. 64, n. 7, p. 1330–1344, 2013. Disponível em: doi: 10.1002/asi.22853. Acesso em: 20 set. 2020.

YANG, L. Metadata effectiveness in internet discovery: an analysis of digital collection metadata elements and internet search engine keywords. **College & Research Libraries**, [s.l.], v. 77, n. 1, p. 7-19, jan. 2016. Disponível em: doi: 10.5860/crl.77.1.7. Acesso em: 18 set. 2020.

ZAVALINA, O. L. Contextual metadata in digital aggregations: application of collection-level subject metadata and its role in user interactions and

information retrieval. **Journal of Library Metadata**, [s.l.], v. 11, n. 3-4, p. 104-128, 2011a. Disponível em: doi: 10.1080/19386389.2011.629957. Acesso em: 25 set. 2020.

ZAVALINA, O. L. Free-text collection-level subject metadata in large-scale digital libraries: a comparative content analysis. *In*: DCMI INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2011. **Proceedings** [...]. 2011b. p. 147-157.

ZHANG, J.; YU, Q.; ZHENG, F.; LONG, C.; LU, Z.; DUAN, Z. Comparing keywords plus of WOS and author keywords: a case study of patient adherence research**.** Journal **of the for Information Science and Technology**, [s.l.], v. 67, n. 4, p. 967–972, 2016. Disponível em: doi: 10.1002/asi.23437. Acesso em: 18 set. 2020.

ANALYSIS OF KEYWORDS OF SCIENTIFIC PRODUCTION OF RESEARCHERS: THE AUTHOR AS AN INDEXER

ABSTRACT

Introduction: With the advent of the digital and web era, the keyword has become an essential representation product in open access information storage and retrieval systems. It is used in the extraction functions for the purpose of scientific identification of authors, bibliometric analysis, indicators of scientific impact, development of controlled vocabularies and other systems of knowledge organization. The attribution of keywords by the author in scientific publications is a practice of representing the content carried out when filling in metadata. This assignment goes through personalized subject analysis and depends on the author's specialty vocabulary, which, in general, has no guidance on standardization and vocabulary control. On the other hand, such subject metadata that receives the keyword does not undergo professional validation. **Objective:** Analysis of keywords assigned by researchers to submit articles from journals indexed in Scopus and in the Portal Professores Unesp regarding the standardization and vocabulary control for different functions in information storage and retrieval systems. **Methodology:** For this purpose, an exploratory research was carried out with an observation study and analysis of keywords attributed by researchers from Portal Docentes Unesp based on data from the CNPq Lattes Curriculum compared to keywords attributed to journal articles. **Results:** The results demonstrate an absence of standardization in the keywords of the journal articles of researchers registered in the Lattes Curriculum, both at syntactic and semantic level. Regarding the assessment of indexation, low levels of consistency are observed when compared to the original articles, both in the rigid index and in the relaxed / flexible index. **Conclusions:** There is a need to develop a policy for organizing and representing information that provides guidelines to researchers regarding the attribution of keywords, aiming at greater standardization and consistency both in the representation and in the recovery of their scientific production.

Descriptors: Keyword. Indexing. Scientific production. Control of vocabulary. Controlled vocabulary. Lattes Curriculum

ANÁLISIS DE PALABRAS CLAVE DE PRODUCCIÓN CIENTÍFICA DE INVESTIGADORES: EL AUTOR COMO INDIZADOR

RESUMEN

Introducción: Con el advenimiento de la era digital y web, la palabra clave se ha convertido en un producto de representación esencial en los sistemas de almacenamiento y recuperación de información de acceso abierto. Se utiliza en las funciones de extracción con fines de identificación científica de autores, análisis bibliométrico, indicadores de impacto científico, desarrollo de vocabularios controlados y otros sistemas de organización del conocimiento. La atribución de palabras clave por parte del autor en publicaciones científicas es una práctica de representación del contenido realizada al cumplimentar metadatos. Esta tarea pasa por un análisis personalizado de la materia y depende del vocabulario de especialidad del autor, que, en general, no tiene orientación sobre estandarización y control de vocabulario. Por otro lado, los metadatos del tema que reciben la palabra clave no se someten a validación profesional. **Objetivo:** Análisis de las palabras clave asignadas por los investigadores para enviar artículos de revistas indexadas en Scopus y en el Portal Professores Unesp sobre la estandarización y control de vocabulario para diferentes funciones en sistemas de almacenamiento y recuperación de información. **Metodología:** Para ello, se realizó una investigación exploratoria con estudio de observación y análisis de palabras clave atribuidas por investigadores del Portal Docentes Unesp a partir de datos del Currículo CNPq Lattes en comparación con palabras clave atribuidas a artículos de revistas. **Resultados:** Los resultados demuestran una ausencia de estandarización en las palabras clave de los artículos de revistas de investigadores registrados en el Currículo Lattes, tanto a nivel sintáctico como semántico. En cuanto a la valoración de la indexación, se observan bajos niveles de consistencia respecto a los artículos originales, tanto en el índice rígido como en el índice relajado / flexible. **Conclusiones:** Existe la necesidad de desarrollar una política de organización y representación de la información que brinde pautas a los investigadores en cuanto a la atribución de palabras clave, buscando una mayor estandarización y consistencia tanto en la representación como en la recuperación de su producción científica.

Descriptores: Palabra clave. Indexación. Producción científica. Control de vocabulario. Vocabulario controlado. Plan de estudios de Lattes.

Recebido em: 15/09/2020

Aceito em: 30/09/2020