O uso da folksonomia na atualização de vocabulários controlados da área da Pediatria[1]

The use of folksonomy in the updating of controlled vocabulary in the area of Pediatrics

Luciana Monteiro Krebs

Universidade Federal do Rio Grande do Sul, Brasil
luciana.monteiro@ufrgs.br

Rita do Carmo Ferreira Laipelt

Universidade Federal do Rio Grande do Sul, Brasil
rita.laipelt@ufrgs.br

Samuel Santos da Rosa

Universidade Federal do Rio Grande do Sul, Brasil
samuel.sdrosa@ufrgs.br

Resumo

Apresenta análise transversal de folksonomias para estudo da linguagem dos usuários de informação na área de Pediatria. Tem como pressuposto de pesquisa que as *tags* podem ser fonte de coleta de candidatos a termos para atualização de vocabulários controlados. O objeto de estudo são as *tags* armazenadas e compartilhadas através do ambiente online CiteULike. Usou-se dois conjuntos de dados, cada um formado por um *corpus* de pesquisa (*tags*) e um *corpus* textual (artigos científicos compartilhados por pelo menos dois usuários da ferramenta). As *tags* foram validadas tanto no *corpus* textual quanto em um vocabulário controlado da área médica. Foram analisadas 195 *tags* de 2011, das quais 93% foram validadas, e 282 *tags* de 2017, com validação de 79%. Conclui-se que as *tags* são boas fontes para a identificação de candidatos a termos, assim como de variantes para enriquecer vocabulários controlados e alimentar sistemas de remissivas de catálogos de autoridades de assunto.

Abstract

The paper presents a transversal analysis of folksonomies to investigate the language of information users in the area of Pediatrics. It has as research hypothesis the idea that the tags can be a source for collecting term candidates to update controlled vocabularies. The object of study is the tags stored and shared through the CiteULike online platform. Two sets of data were used, each consisting of a research corpus (tags) and a textual corpus (scientific papers shared by at least two tool's users). The tags were validated both in the textual corpus and in a controlled vocabulary of the medical area. We analyzed 195 tags from 2011, of which 93% were validated, and 282 tags from 2017, with validation of 79%. It is concluded that tags are good sources for identifying term candidates as well as variants for enriching controlled vocabularies and feeding referral systems from subject authority catalogs.

Palavras-chave: Representação da informação. Organização do conhecimento. Vocabulários controlados. Folksonomi.

Keywords: *Information representation. Knowledge organization. Controlled vocabularies. Folksonomy.*

1. Introdução

Em 2007, 94% das informações produzidas no mundo foram armazenadas em formato digital (Hilbert & López, 2011). Uma previsão da Cisco (2013) indica que em 2017 serão produzidos 7,7 zetabytes de dados. Enquanto fazemos uso de recursos tecnológicos que possibilitam atravessar fronteiras, buscar conhecimento, relacionamentos e experiências que talvez pessoalmente não possamos realizar, tudo isso fica registrado, gerando *logs* de dados em grandes volumes, e muitas vezes acessível a públicos que o próprio autor desconhece. Em várias plataformas é possível compartilhar deliberadamente o registro destas atividades *online* e, na maioria dos casos, é o que se deseja, o que revela uma mudança social relevante em termos de privacidade. Além disto, a informação é registrada em escalas tão grandes, que encontrar uma informação que se busca pode se tornar uma tarefa demasiadamente complicada.

O aumento exponencial de informação disponibilizada através da internet sinaliza para uma expansão significativa de competências, cujos profissionais da informação devem se apropriar se quiserem continuar exercendo um papel importante para os usuários. *Blogs*, fóruns de discussão e até redes sociais podem ser fontes de informação para novos usuários, e estes recursos não só cresceram em quantidade, mas revelam-se complexos em termos tecnológicos, que também avançam todos os dias. Os processos de representação e recuperação da informação tornam-se cada vez mais sofisticados, com o incremento de inteligência artificial nos motores de busca, por exemplo. Com o advento de tecnologias para compartilhamento de informações e meta-informações através da web, podemos estar diante de uma significativa oportunidade de melhoria no processo de indexação, contando com a contribuição de metadados fornecidos pelos próprios usuários em linguagem natural.

A presente pesquisa visa refletir sobre a possível contribuição da folksonomia nos processos formais de indexação de documentos (especificamente na representação temática) enquanto fontes de coleta de candidatos a termos, objetivando aproximar a linguagem utilizada nestes

processos daquela conhecida pelo usuário. Almeja-se com isso, alcançar maior eficácia e assertividade no momento da busca, poupando o tempo do leitor.

A utilização do CiteULike como plataforma para coleta de dados justifica-se em função de ser um ambiente exclusivo para compartilhamento de literatura científica. Ele é utilizado, portanto, por um público altamente especializado. Diferente da plataforma social ResearchGate, o CiteULike permite a atribuição de *tags* aos conteúdos compartilhados (a rede social acadêmica previamente mencionada permite atribuição de *tags* apenas como "temas de interesse" do pesquisador). Já o Academia.edu permite a associação de "*bookmarks*" a um artigo, porém apenas quem fez o *upload* – provavelmente seu autor – pode atribuir as *tags/bookmarks* ao documento, não sendo o mesmo possível para os demais usuários.

2. Folksonomias e a indexação: possíveis aproximações

O desafio de quem procura conhecimento atualmente, não é mais o acesso e permissão para ler determinado documento como ocorria na era medieval. Hoje, o maior desafio de leitores e pesquisadores está em sua capacidade de encontrar, entre tantas publicações, aquela que realmente tem relevância para suprir uma necessidade informacional. Tornou-se mister então, encontrar, de fato, a informação precisa no tempo adequado. Agilidade e precisão na recuperação da informação têm ainda mais valor no cenário atual com os recursos tecnológicos disponíveis e a descentralização de entidades produtoras de informação propiciadas pela internet.

Nesse contexto, a realidade dos usuários que buscam por informação também se modificou significativamente, com a possibilidade de acessarem o conteúdo de catálogos e bases de dados de qualquer unidade de informação à distância. Consequentemente, o bibliotecário passa a ter de lidar com outros recursos para tentar dar conta de "conhecer" este usuário que raramente vê pessoalmente, considerando alternativas de aproximação com esses usuários que se encontram cada vez mais distantes fisicamente.

Essa realidade exige um esforço redobrado por parte dos profissionais, para manter a unidade de informação, e o seu próprio trabalho, relevantes. É vital aprender como a internet funciona, como as pessoas se comunicam através dela, e o que se pode aprender dali. O conceito em si não é novo. A importância de aproximar o máximo possível a linguagem de indexação do

repertório da comunidade usuária já foi sinalizada há muito tempo. A literatura da Ciência da Informação, frequentemente, preconiza a aproximação da linguagem utilizada na indexação com a linguagem natural do usuário, para que este obtenha, cada vez mais, sucesso nos seus objetivos de busca.

Em 1876, Cutter, em sua obra "*Rules for a dictionary catalog*" determina regras para a formação de cabeçalhos alfabéticos de assuntos[2]. Fujita, Rubi & Boccato (2009, p. 26) resumem as regras em três princípios básicos, entre os quais se chama atenção para o primeiro: "Princípio do uso: as descrições devem ser feitas da forma usada pelo usuário". Gomes, Motta & Campos (2006, *online*) encontram na quarta lei de Ranganathan (Poupe o tempo do leitor) uma premissa para o tratamento dos termos no processo de indexação:

> [...] dois dispositivos contribuem para o atendimento a esta Quarta Lei:
>
> 1) adoção do termo mais plausível de ser buscado pelo usuário, quanto ao uso corrente e quanto à especificidade;
>
> 2) adoção de um dispositivo no sistema de recuperação de informação de sorte que, mesmo buscando por um termo não-preferido, o sistema automaticamente aceite o termo e recupere a informação via termo preferido. Assim, com um único passo, o leitor acessaria a base de dados. (Gomes, Motta & Campos, 2006, online).

O termo não-preferido pode ser entendido aqui como as variações terminológicas, variações do termo escolhido, ou seja, variantes podem tornar-se pontos de acesso no catálogo, sem destituir o termo escolhido como descritor. Assim, o bibliotecário deve aproximar-se do usuário tanto quanto possível, e dentro do ambiente específico da internet, isto significa compreender o uso das ferramentas utilizadas para se comunicar e também para representar a informação, como as *tags*.

Essa aproximação é importante, sobretudo, para impedir que a informação fique "oculta" dentro dos acervos – devido a processos de indexação pouco eficazes. Por isso, é recomendável que sejam realizados estudos de comunidades e usuários para compreender sua linguagem e incluí-la nos sistemas de recuperação da informação através de remissivas. Sendo os sistemas eletrônicos de informação binários, a simples ausência de uma remissiva dentro do conjunto de termos presentes no sistema pode apresentar um resultado vazio para a busca, mesmo que o documento desejado esteja no acervo. Espera-se do bibliotecário que realize esse filtro sofisticado, para agregar cada vez mais formas de tornar a busca do usuário eficaz.

A folksonomia pode ser entendida como uma representação simples de conteúdos na internet (páginas, *links*, fotografias, textos, músicas, vídeos, documentos em geral), que reforça laços sociais e expressa linguagem natural dos usuários desta informação. O termo foi cunhado por Thomas Vander Wal, arquiteto da informação, que define a etiquetagem atribuída a conteúdos na *web* pelos usuários, com o objetivo de recuperar este conteúdo em um dado momento. Segundo Vander Wal (2007), folksonomia é o resultado da marcação livre e pessoal de informações e objetos (qualquer coisa com uma URL) para uma recuperação própria [tradução nossa]. Neste caso, a marcação a que o autor se refere são as etiquetas (ou *tags*), atribuídas pela pessoa que consome a informação. A marcação é feita em um ambiente social (compartilhado e aberto aos outros).

Normalmente, nos sistemas *online* de compartilhamento de documentos (artigos, textos, fotos, vídeos, etc.) é possível, além de atribuir as *tags* para representar a informação, recuperar documentos através da *tag* (quando esta é transformada em *link)*. Este *link*, ao ser clicado, retorna uma lista de documentos aos quais a *tag* foi atribuída, permitindo que o usuário visualize e continue sua navegação através desta forma de busca. Para Amaral e Aquino (2008), no momento em que a prática das *tags* permite a qualquer usuário representar e recuperar informações através de etiquetas criadas livremente e com base nos significados dos dados etiquetados, ela surge como uma alternativa de gerenciamento de informação.

Dentro do processamento técnico de uma unidade de informação, o estabelecimento dos descritores dos itens da coleção é uma das atividades mais importantes, pois é o que garante a recuperação - ou não - dos documentos pelos usuários. Esta etapa deve ser feita de tal forma que garanta precisão evitando equívocos e inconsistências no resultado apresentado ao leitor, e que fica disponível no catálogo. O quadro 1 tem a função de delimitar as diferenças entre folksonomias e vocabulários controlados, pois ambos são utilizados nesta pesquisa.

Quadro 1 – ***Tabela comparativa de folksonomias e vocabulários controlados***

	Folksonomias	**Vocabulários controlados**
Objetivo	Representar e recuperar a informação.	Representar e recuperar a informação.
Plataforma de registro	Exclusivamente na web.	Sistemas de organização do conhecimento (catálogos, bibliotecas), sejam online ou off-line.
Autoria da classificação	Usuários da informação e autores.	Bibliotecários e profissionais especializados.
Linguagem	Natural.	Artificial.
Momento em que o material é classificado	Durante ou após o uso da informação.	Antes de disponibilizar a informação para o uso.

Fonte: os autores

Como se pode observar, há diferenças entre as folksonomias e os vocabulários controlados, no que tange à plataforma de registro, autoria da classificação, e outros, porém, ambas são utilizadas com o objetivo de representar e recuperar a informação.

Levanta-se, portanto, a questão sobre se a informação disponível em forma de *tags* pode ser aproveitada em estudos e eventualmente utilizada. Especialmente em relação às expressões empregadas pelos usuários na representação do conhecimento - e que fazem parte do repertório pessoal individual destes - como contributo para enriquecer sistemas de catalogação de documentos enquanto recursos para aproximar o usuário da informação que ele necessita ou deseja.

3. Metodologia

Essa seção apresenta as etapas metodológicas realizadas na pesquisa. Os resultados apresentados são fruto de um estudo comparativo entre conjuntos de dados, sendo um de 2011 e outro de 2017. Cada um dos conjuntos foi formado por um *corpus* de pesquisa (*tags* atribuídas pelos usuários) e um *corpus* textual (texto completo dos artigos para validação). Foi utilizado também, para validação, um vocabulário controlado da área médica. Optou-se por utilizar uma ferramenta de armazenamento e compartilhamento *online* de documentos com o recurso de *social tagging*, o CiteULike[3], para avaliar os resultados de atribuição de etiquetas por usuários nos conteúdos, de forma que se pudesse avaliar a relação entre estas etiquetas e termos de vocabulários controlados, além de identificar sua presença ou não dentro do *corpus* textual.

3.1. Planejamento e critérios de busca

A composição dos conjuntos de dados ocorreu de acordo com as seguintes diretrizes: a) foram admitidos artigos científicos a que se obteve acesso ao texto integral; b) artigos categorizados pela *tag* "*pediatrics*" por qualquer usuário da ferramenta CiteULike e em qualquer data; c) artigos compartilhados por pelo menos dois usuários do ambiente.

Nesta pesquisa optou-se por utilizar a busca simples, que inclui todos os artigos públicos e autenticados no CiteULike. Sabe-se que a busca simples no CiteULike não leva em consideração apenas as *tags* atribuídas pelos usuários. Se fosse assim, teríamos na lista de *tags* tantas ocorrências da *tag* "*pediatrics*" quantos documentos recuperados no *corpus*, o que não ocorre.

Também foram analisadas as *tags* atribuídas a cada documento, e ocorre que vários documentos não receberam a *tag* "*pediatrics*" de nenhum usuário. Assim, podemos inferir que a ferramenta de busca simples do CiteULike leva em consideração também outras informações que não somente *tags*, provavelmente título do documento, nome do periódico em que o artigo foi publicado, informações do texto ou resumo etc. Atualmente é possível fazer uma busca avançada na ferramenta apenas pelas *tags* atribuídas aos documentos (usando a expressão *tag:pediatrics*), no entanto, como o objetivo desse trabalho é realizar uma análise transversal, comparando os dados de 2011 e 2017, optou-se por manter o uso da busca simples.

A primeira coleta ocorreu em 27 de agosto de 2011. A segunda em 8 de agosto de 2017. Na página inicial do CiteULike, digitou-se a palavra-chave "*pediatrics*" no campo "*Search citeulike*" no canto direito superior da tela. Os resultados são listados aparentemente sem ordenamento - nem cronológico, nem alfabético, nem por relevância. A seguir estão detalhadas as etapas de coleta, limpeza e organização de cada um dos *corpora*.

3.2. Coleta, limpeza e organização do *corpus* textual

Em relação ao *corpus* textual, foram recuperados 903 documentos em 2011 e 850 em 2017, dos quais foram selecionados apenas os compartilhados com dois ou mais membros, o que ocorreu em 177 artigos em 2011, e 190 em 2017. Copiaram-se as seguintes informações de

cada artigo (metadados fornecidos pelo CiteULike): título, referência bibliográfica, autor, *tags* atribuídas ao documento, número de pessoas que compartilham o documento.

Foi realizado o *download* de todos os artigos encontrados no resultado de busca que atendiam aos critérios estabelecidos, com exceção dos que, por alguma restrição de acesso, não se pôde realizar o *download*. Ao pesquisar nas bases de dados disponíveis no Portal de Periódicos da Capes[4] e no Google Acadêmico[5], tivemos acesso a **156** dos 177 artigos inicialmente selecionados em 2011 e **185**[6] dos 190 inicialmente selecionados em 2017. Todos os documentos se encontram em língua inglesa. Cada documento recebeu um número.

Os arquivos foram então convertidos do formato .PDF para .TXT para que pudessem ser mais facilmente analisados e lidos pelas ferramentas de processamento de *corpus*. Foram retiradas dos arquivos de texto todas as partes dos artigos não necessárias para a posterior análise terminológica, como as tabelas, referências bibliográficas, ilustrações, datas, notas de rodapé, currículo dos autores e paginações, para que permanecesse no arquivo apenas o artigo em si.

Na próxima etapa realizou-se o *upload* do *corpus* textual na ferramenta e-Termos[7]. O e-Termos é um ambiente colaborativo *online* de acesso gratuito cujo objetivo é auxiliar na gestão terminológica. Foi cadastrado um projeto com o nome Pediatria, feito *upload* dos textos (Gênero "Científico" e Tipo Textual "Artigo") e em seguida foi feita a compilação dos *corpora*.

3.3. Coleta, limpeza e organização do *corpus* de pesquisa

Para coletar o *corpus* de pesquisa (*tags*), clicou-se no registro de cada documento, copiou-se o trecho "*posting history*" e colaram-se todas estas informações em um arquivo de texto. Em *posting history* são exibidas as informações do histórico de compartilhamento, ou seja, data, *nickname* do usuário e *tags* atribuídas por cada um deles no momento do compartilhamento. Os conjuntos de *tags* por artigo receberam a numeração definida para o artigo correspondente.

A atribuição de *tags* no CiteULike ocorre quando o usuário compartilha o documento. Um campo de texto livre é usado para este fim, e as *tags* são separadas por espaço simples, ou seja, a única forma de usar expressões como *tag* é introduzindo um sinal gráfico como underline (_) ou hífen (-) ou ainda escrevendo os dois termos sem separação de espaço,

porque juntos representam um sintagma. São exemplos: *intervention_services*, *alcohol-abuse* e *familyphysician* (para o conceito "médico de família"). Enquanto alguns usuários utilizam o *underline*, outros preferem escrever as expressões sem o espaço, e outros ainda ignoram esta limitação da ferramenta e mantém os espaços, o que separa os termos que compõem uma expressão na hora de analisar as *tags*. Assim, alguns termos dentro das *tags* levantadas ficam sem sentido devido à quebra do sintagma terminológico, e por este motivo foram retirados do *corpus* de pesquisa.

Em relação ao *corpus* de pesquisa de 2011, de um total de 979 *tags* atribuídas aos documentos, foram retiradas as *tags* repetidas e restaram 493 *tags* únicas. Já no *corpus* de 2017, de um total de 1.494 *tags*, 651 eram *tags* únicas. A limpeza das *tags* serviu para eliminar expressões com erros de grafia, *tags* que não possuem sentido claro quando empregadas ao assunto Pediatria, ou aquelas que, apesar de estarem no discurso dos especialistas, são muito genéricas e transitam em várias áreas, ou seja, são pouco específicas deste campo. Durante este processo foram retiradas algumas *tags* que possuíam erros de grafia, como "*childrens*" (a grafia correta é *children*) e "*breast_feeding*" (a redação correta é *breastfeeding*).

Também se identificou que algumas *tags* encontravam-se em outros idiomas, como turco, alemão e português e que tinham relação com a temática dos artigos pesquisados, mas foram retiradas em função das ferramentas de validação (*corpus* textual e vocabulário controlado) estarem no idioma inglês, o que traria resultado inválido. Algumas destas ocorrências estão registradas a seguir: a) *cocuk* – em turco, criança; b) *siber_zorbalik* – (*siber zorbalik*) em turco, cyber-bullying; c) *epilepsia* – em português; d) *auswirkungen* – em alemão, efeito.

Finalmente, mesmo percebendo que algumas *tags* poderiam ser variantes, como "*neonatal_thrombocytopenia*" e "*low platelet count*", optou-se por mantê-las no *corpus* de pesquisa para verificar sua ocorrência nas ferramentas de validação. As siglas encontradas no *corpus* de pesquisa de 2017 foram tratadas à parte, uma vez que as mesmas não foram coletadas e analisadas no *corpus* de 2011. Esta medida foi tomada para não distorcer os resultados comparativos.

Após a limpeza do *corpus* de pesquisa, resultaram **195** *tags* a serem analisadas no *corpus* de 2011 e **282** no *corpus* de 2017. Foi então realizada a normalização das *tags*, corrigindo o uso de hífen e *underline* onde deveria haver espaços (no caso de sintagmas), ou mesmo a

supressão do espaço, devido à limitação do CiteULike descrita anteriormente. São exemplos desta normalização: "*alcohol-abuse*" foi substituído por "*alcohol abuse*", "*speech_and_language_delay*" foi substituído por "*speech and language delay*" e "*earlyintervention*" foi substituído por *"early intervention*".

3.4. Validação das *tags*

Nesta etapa realizou-se a validação de ocorrência das *tags* no *corpus* textual através do e-Termos. Verificou-se que, de acordo com a tipologia do termo analisado, é necessário utilizar uma ferramenta diferente em função da maneira como a frequência é calculada em cada uma delas. Por isso, as ferramentas utilizadas foram: "Contador de Frequência" para termos simples, "Consulta Termos" para sintagmas (termos compostos) e "Identificador de Siglas, Acrônimos e Nomes Próprios" para siglas.

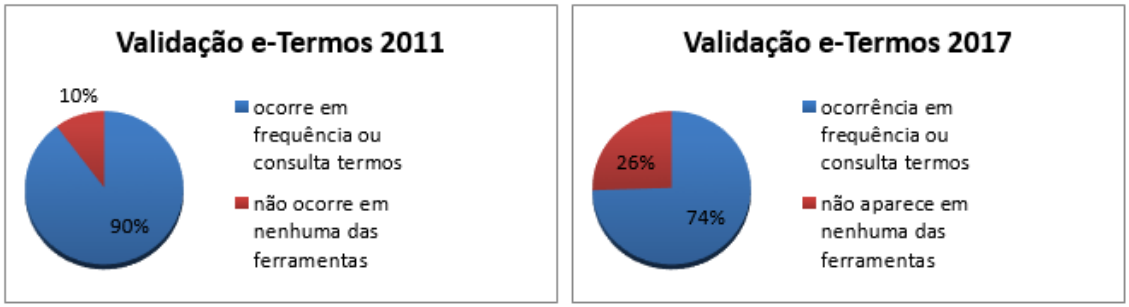
A próxima etapa consistiu na validação das *tags* em vocabulário controlado da área de Medicina. Para isso, utilizou-se a ferramenta Descritores em Ciências da Saúde (DeCS)[8]. Para o resultado do DeCS foram adotadas as seguintes categorias: (a) encontrado: quando a *tag* foi encontrada como descritor com exatamente a mesma grafia; (b) parte de descritor: quando a *tag* foi encontrada como parte de um descritor; (c) variante: quando a *tag* foi encontrada, mas não como descritor e sim variante ou parte de variante no DeCS; e (d) não encontrado: quando a *tag* não pôde ser validada, pois não foi encontrada no DeCS.

A partir do resultado da validação do *corpus* de pesquisa foi possível calcular quanto do *corpus* de pesquisa efetivamente apareceu no *corpus* textual e/ou no vocabulário controlado. A título de ilustração, o **Apêndice A** traz as dez *tags* mais compartilhadas nos *corpora* de 2011 e 2017. A próxima seção refere-se à análise dos dados.

4. Análise e resultados da pesquisa

No *corpus* de pesquisa de 2011, das 195 *tags* selecionadas para análise, 175 foram validadas no *corpus* textual (e-Termos), aparecendo pelo menos uma vez. Já no *corpus* de pesquisa de 2017, das 282 *tags* selecionadas para análise, 210 foram validadas no *corpus* textual (e-Termos). Na Figura 1 pode-se visualizar os percentuais de ocorrência em cada um dos períodos.

Figura 1 – ***Validação comparada das tags no corpus textual***



Fonte: dados da pesquisa

Segundo os dados da pesquisa apresentados na Figura 1, 90% das *tags* de 2011 foram validadas no *corpus* textual, e em 2017 a validação foi de 74%. Na análise transversal, pode-se observar a diferença de 10% para 26% de *tags* que não ocorrem no e-Termos. Isso significa que entre 2011 e 2017 a distância entre a linguagem do usuário (representada nas *tags*) e a linguagem dos especialistas constante nos textos aumentou.

Como a ferramenta onde os documentos são marcados pela inserção de *tags* é de interesse de especialistas, pode-se inferir que esse afastamento se deve a escolhas linguísticas/terminológicas dos usuários. Sabe-se que a linguagem utilizada para a escrita de artigos científicos é mais formal por exigência do próprio gênero textual. A linguagem utilizada para a marcação do conteúdo de documentos em plataformas sociais de compartilhamento, por outro lado, é livre e informal, apresentando em alguns casos *tags* de interesse particular dos usuários que não representam o conteúdo do documento marcado (como as tags "*my_phd*" ou "*literature_search*"). Mesmo que *tags* com essa particularidade tenham sido retiradas dos *corpora* de pesquisa durante a limpeza, percebe-se que o usuário faz uso do seu repertório linguístico individual muitas vezes, pois essa é a característica proeminente da folksonomia. Outra possibilidade para justificar esse afastamento é que as *tags* não encontradas sejam neologismos, e nesse caso é natural que não ocorram no *corpus* textual, por serem termos que ainda estão começando a fazer parte da linguagem especializada no campo, sem terem se consolidado em glossários e vocabulários controlados da área.

Em vocabulários controlados, a ausência de variantes causa dificuldades de recuperação de informação pelos usuários. Quando um documento está indexado com um termo, mas o usuário realiza a busca utilizando outro, se esta variante não estiver registrada a recuperação é prejudicada. Faulstich (2002, p. 70) afirma que os termos, no meio linguístico e social, "são

entidades passíveis de variação e mudança". Entende-se que os profissionais da informação devem estar atentos a estas variações e que seus produtos estejam alinhados com as expressões de busca.

Por isso, além de validar o *corpus* de pesquisa no *corpus* textual, neste trabalho realizou-se a validação das *tags* também no vocabulário controlado. Nesta etapa, reuniram-se as *tags* que apareceram de alguma forma no DeCS, seja como um descritor exatamente com a mesma grafia, seja encontrado como variante ou parte de variante de um descritor ou então como parte de um descritor composto por mais de uma palavra. A Figura 2 representa os resultados obtidos em 2011 e 2017.

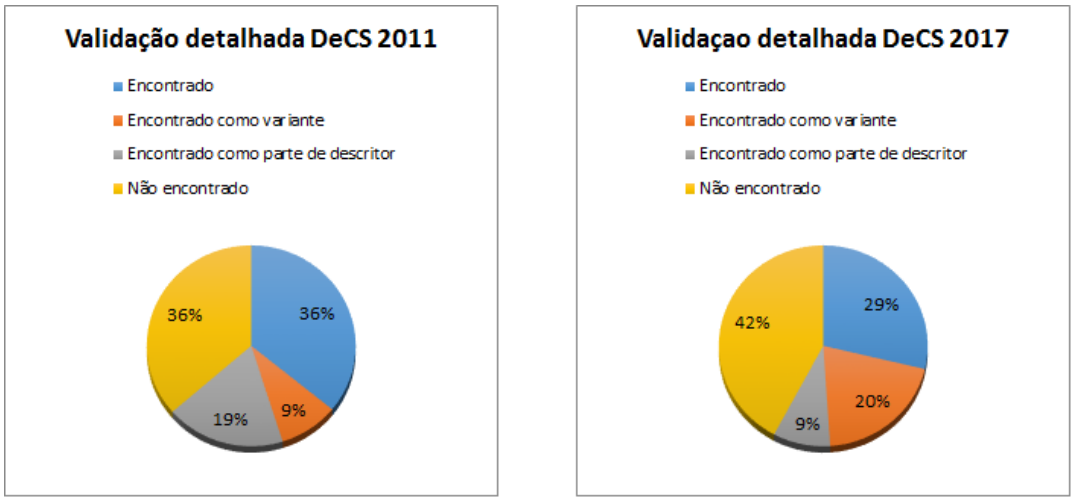
Figura 2 – ***Validação comparada das tags no vocabulário controlado***

Fonte: dados da pesquisa

Na análise transversal pode-se observar a diferença de 36% para 42% de *tags* que não ocorrem no DeCS. Isso significa que entre 2011 e 2017 a distância entre a linguagem do usuário (representada nas *tags*) e a linguagem controlada (representada pelos descritores e variantes do DeCS) também aumentou, assim como ocorreu com a validação das *tags* no *corpus* textual. Como os descritores do DeCS são utilizados por bibliotecários para a indexação de obras da área da saúde, pode-se dizer então que observamos um maior afastamento entre a linguagem dos usuários e a linguagem usada para indexação pelos bibliotecários da área da saúde.

Por outro lado, embora tenhamos identificado esse distanciamento, percebe-se, também, que a presença de variantes terminológicas, encontradas no DeCS em 2017, aumentou em relação às variantes encontradas em 2011, indo de 9% para 20%. Esse resultado é positivo, pois pode indicar que os gestores do DeCs estão cientes da importância das variantes para a recuperação da informação. A Figura 3 apresenta o detalhamento da validação realizada, na qual é possível identificar como as *tags* foram encontradas (se como descritores ou variantes).

Figura 3 – ***Validação comparada das tags no vocabulário controlado***

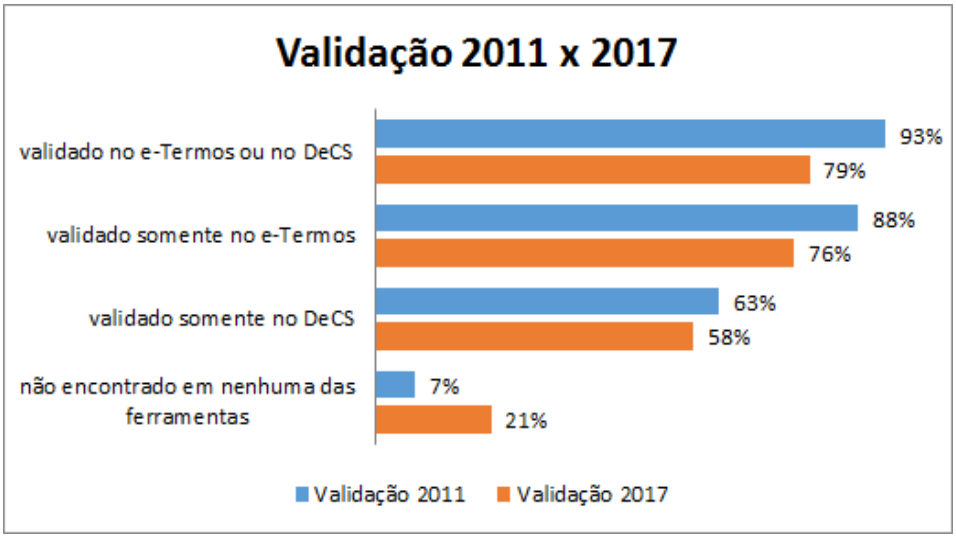


Fonte: dados da pesquisa

Outro aspecto importante é que nenhuma das variáveis (*corpus* textual ou DeCS) **isoladamente** é responsável pelo distanciamento geral da linguagem do usuário em relação à linguagem expressa nos textos (Figura 1) e à linguagem do vocabulário controlado (Figura 2). O percentual de *tags* não encontradas poderia ter sido o mesmo se apenas uma das linguagens (do especialista ou do bibliotecário) estivesse muito distante das *tags* atribuídas, mas isso não ocorre. Tanto a validação no e-Termos (que apresentou 88% em 2011 e 76% em 2017) quanto a validação no DeCS (que apresentou 63% em 2011 e 58% em 2017) contribuem para o aumento do percentual de *tags* não validadas em 2017.

Pode-se verificar, na Figura 4, que a diferença de validação entre 2011 e 2017 foi de 12% no *corpus* textual (e-Termos) e 5% no DeCS (vocabulário controlado) respectivamente.

Figura 4 – ***Validação 2011 x 2017***



Fonte: dados da pesquisa

Para entender as possíveis causas desse resultado, analisaram-se as *tags* do *corpus* de pesquisa que não foram validadas em nenhuma das ferramentas. Estas *tags*, que não aparecem nem no discurso dos especialistas (validação no *corpus* textual através do e-Termos) tampouco no vocabulário controlado (validação através do DeCS), somam 14 no *corpus* de 2011 e 59 no *corpus* de 2017. Na Tabela 1 encontram-se as 10 *tags* não validadas com maior frequência no CiteULike.

Tabela 1 – ***Tabela de tags sem ocorrência no corpus textual ou no DeCS***

2011		***2017***	
***Tag* normalizada**	**Freq. CiteULike**	***Tag* normalizada**	**Freq. CiteULike**
neurological impairments	3	autism asd	24
Pulm	3	attention deficit hyperactivity disorder adhd	18
abuse teen	2	neurological impairments	5
blood brain barrier	2	nicu nbn	4
constraint induced movement therapy	2	otitis media with effusion ome	4
genetic screening newborn	2	measles mumps rubella mmr vaccine	3
Hypoalbuminaemia	2	abuse teen	2
neurology physical therapy	2	development disorder pervasive	2
autism asd	1	developmental coordination disorder dcd	2
infeccion urinaria	1	gluten free casein free diet	2
Lactância	1	joint hypermobility syndrome jhs	2

Fonte: dados da pesquisa.

Nessa análise observamos que, apesar de não validadas no e-Termos nem no DeCS, essas *tags* não estão desconectadas do contexto da Pediatria, o que indica a possibilidade de serem novos termos (novos demais para já terem sido registrados em vocabulários controlados). A comparação constante de *tags* em relação a descritores e variantes confere dinamismo ao sistema de recuperação da informação e alinha buscas e resultados.

No *corpus* de 2017, percebeu-se também outra característica. Aparentemente, o motivo pelo qual a maior parte das *tags* não foi validada foi o uso concomitante de termo e sigla, como nas *tags* "*developmental coordination disorder dcd*", "*joint hypermobility syndrome jhs*" e "*attention deficit hyperactivity disorder adhd"*. Ao analisar estas ocorrências pode-se afirmar

que, apesar de serem possíveis bons representantes do conteúdo dos artigos, esta forma de representação dos usuários impede sua validação, pois se as siglas tivessem sido registradas separadamente do termo composto, tanto o termo quanto a sigla teriam sido validados.

Conforme já dito anteriormente, as siglas não foram analisadas no *corpus* de 2011 e por isso não foram incluídas na análise transversal aqui apresentada (juntamente com o *corpus* de 2017). No entanto, na coleta de 2017 identificamos um uso significativo de siglas para representação dos conteúdos dos artigos científicos compartilhados no CiteULike (46 *tags*), sendo exemplos *"adhd"* (frequência 35 CiteULike), *"cpoe"* (12 ocorrências) e *"ome"* (7 ocorrências).

Verificamos que estas siglas se referem a termos da área pediátrica e, analisando-as separadamente, validamos 28 delas no *corpus* textual e 14 no vocabulário controlado (12 como variante ou parte de variante; 1 como parte de descritor e 1 encontrado como descritor). Desta análise pode-se concluir que as siglas são muito usadas pelos usuários e especialistas e por isso, é importante que sejam inseridas pelos bibliotecários como remissivas nos vocabulários controlados.

Adicionalmente, observou-se que alguns conceitos foram representados pelos usuários com variação de numeral, como, por exemplo, "*newborn* / *newborns*" "*immunization* / *immunizations".* Em alguns casos essa variação foi validada, em outros não. Esta é uma característica da folksonomia, com a atribuição livre de etiquetas, que difere da linguagem normalizada dos vocabulários controlados.

5. Conclusão

A partir das análises apresentadas no decorrer desse artigo, conclui-se que as *tags* atribuídas pelos usuários do CiteULike a documentos compartilhados no ambiente podem contribuir, enquanto fontes de coleta de candidatos a termos, para o enriquecimento da representação temática de documentos, seja através das remissivas em vocabulários controlados ou em catálogos de autoridades.

A análise transversal dos *corpora* (2011 e 2017) indica que houve distanciamento ao longo do tempo entre a linguagem dos usuários, especialistas e bibliotecários, pois houve um decréscimo tanto na quantidade de *tags* validadas no DeCS quanto no *corpus* textual,

comparando-se 2011 e 2017. Apesar disto, observou-se um crescimento no percentual de variantes encontradas no DeCS de 2011 para 2017, o que indica um esforço de aproximação com a linguagem dos usuários por parte dos bibliotecários responsáveis pelo vocabulário controlado.

Conclui-se, ainda, que as *tags* que não apareceram no *corpus* textual ou no vocabulário controlado podem ser neologismos, e, portanto, devem ser observadas ao longo do tempo na literatura especializada devido à possibilidade de virem a se tornar termos.

Considerou-se alto o número de siglas encontradas entre as *tags* em 2017, e por este motivo, sugere-se que as mesmas sejam incluídas nos vocabulários controlados como remissivas, para melhorar a recuperação da informação. Para pesquisas futuras, sugere-se que novos trabalhos sejam realizados para análise da linguagem dos usuários, incluindo ambientes de compartilhamento de literatura científica como o próprio CiteULike e o Academia.edu.

Observa-se, por fim, que uma limitação do estudo deve-se a uma característica da ferramenta CiteULike, que separa as *tags* pelo espaço e, portanto, não admite termos compostos como etiquetas. Por este motivo, muitas *tags* precisaram ser retiradas dos *corpora* de pesquisa, pois perdem o sentido quando separadas durante a contagem de frequência.

Referências Bibliográficas

AMARAL, A., & AQUINO, M. C. (2008). Práticas de folksonomia e social tagging no Last.fm. In: Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais, 8, 2008. *Anais...* Paraná: PUC. Recuperado de http://www.din.uem.br/gsii/downloads/waihcws/Praticas-Folksonomia-Social-TaggingLastfm.pdf.

CISCO (2013). *Índice mundial sobre entornos de nube de Cisco: previsión y metodología, 2012–2017*. San José: Cisco. Recuperado de http://docplayer.es/1754500-Indice-mundial-sobre-entornos-de-nube-de-cisco-prevision-y-metodologia-2012-2017.html

FAULSTICH, E. [1998] Termo e variação: tendências no português do Brasil. In: *Socioterminologia.* (Excerto, parte II). Brasília: UnB.

FUJITA, M. S. L., RUBI, M. P. & BOCCATO, V. R. C. (2009) As diferentes perspectivas teóricas e metodológicas sobre indexação e catalogação de assuntos. In: Fujita, M. S. L., Boccato, V. R. C., Rubi, M. P. & Gonçalves, M. C. *A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias.* Um estudo de observação do contexto sociocognitivo com protocolos verbais. São Paulo: SciELO - Editora Unesp. Recuperado de https://books.google.com.br/books?id=KAyV8MVAr8YC&

GOMES, H. E., MOTTA, D. F. & CAMPOS, M. L. A. (2006). Princípios normativos. In: _____. *Revisitando Ranganathan: a classificação na rede.* Rio de Janeiro. Recuperado de http://www.conexaorio.com/biti/revisitando/revisitando.htm

HILBERT, M., & LÓPEZ, P. (2011). The world's technological capacity to store, communicate, and compute information. science, 332(6025), 60-65. Recuperado de http://science.sciencemag.org/content/332/6025/60

VANDER WAL, T. (2007) *Folksonomy definition and Wikipedia.* Recuperado de http://www.vanderwal.net/random/entrysel.php?blog=1750

VAN DER LAAN, R. H., FERREIRA, G. I. S., BONOTTO, M. E. K. K.; NEVES, I. C. B., & GASPERIN, I. M. (2004). *Avaliação de descritores relativos às ciências da informação: relato de pesquisa.* Em Questão, Porto Alegre, 10(2), 337-347.

6. APÊNDICE A - Tags mais compartilhadas

Tabela 1 – ***Dez tags mais compartilhadas no corpus de pesquisa de 2011 (CiteULike)***

Termo	Freq. CiteULike	Freq. e-Termos	Ferramenta utilizada	Resultado DeCS
children	16	4391	*frequência*	encontrado como parte de descritor
asthma	12	558	*frequência*	encontrado
neonatal	10	148	*frequência*	encontrado como parte de descritor
adolescent	8	199	*frequência*	encontrado
analgesia	8	48	*frequência*	encontrado
prevention	8	199	*frequência*	encontrado como parte de descritor
childhood	6	296	*frequência*	encontrado como parte de descritor
cognitive	6	231	*frequência*	encontrado como parte de descritor
visual motor	6	12	*consulta termos*	não encontrado
child	5	1269	*frequência*	encontrado

Fonte: dados da pesquisa.

Tabela 2 – ***Dez tags mais compartilhadas no corpus de pesquisa de 2017 (CiteULike)***

Termo	Freq. CiteULike	Freq. E-Termos	Ferramenta utilizada	Resultado DeCS
children	24	5436	*frequência*	encontrado como variante
autism asd	24	0	*consulta termo*	não encontrado
autism	22	389	*frequência*	encontrado como variante
attention deficit hyperactivity disorder adhd	18	0	*consulta termo*	não encontrado
early intervention	12	54	*consulta termo*	encontrado como variante
child	10	1783	*frequência*	encontrado
screening	9	981	*frequência*	encontrado como variante
epidemiology	8	28	*frequência*	encontrado
abuse	8	329	*frequência*	encontrado como parte de descritor
medical	8	691	*frequência*	encontrado como parte de descritor

Fonte: dados da pesquisa

Notas

[1] O presente trabalho foi realizado com o apoio da Pró-Reitoria de Pesquisa - UFRGS – Brasil.

[2] CUTTER, Charles A. Rules for a dictionary catalogue. Washington: Government Printing Office, 1889. Disponível em < http://babel.hathitrust.org/cgi/pt?id=wu.89101448975 >. Acesso em 25 jul. 2013.

[3] http://www.citeulike.com/

4 http://www.periodicos.capes.gov.br/

5 http://scholar.google.com.br/

6 A título de informação, destes 185 artigos, 71 já constavam no corpus textual de 2011.

7 http://www.etermos.cnptia.embrapa.br/

[8] http://decs.bvs.br/