# LINGUAGEM NATURAL OU LINGUAGEM CONTROLADA?[1]
## influência da palavra-chave na representação para indexação e recuperação de informações

# NATURAL LANGUAGE OR CONTROLLED LANGUAGE?
## The influence of the keyword in the representation for indexing and searching for information

*Mariangela Spotti Lopes Fujita*[2]

**RESUMO**
Trata-se de ensaio para apresentar o debate entre linguagem natural ou linguagem controlada na perspectiva contemporânea da internet e web 2.0 em que as atuais bases de dados, muitas de acesso aberto como os repositórios, são heterogêneas quanto à organização e representação da informação. O objetivo é investigar, por meio de revisão de literatura, o uso de linguagem natural ou de linguagem controlada por sistemas de informação, de modo geral, e por repositórios, de modo mais específico, para observar, avaliar e discutir os resultados, recomendações e reflexões. Na análise qualitativa foram selecionados 90 trabalhos que, após leitura prévia, foram escolhidas 73 produções científicas. Na sequência, a leitura foi extensiva para coleta dos objetivos, metodologia, resultados, conclusões e principais pontos de vista dos autores. Uma amostra representativa de 38 textos está distribuída em 12 categorias de análise para facilitar a compreensão macro e micro desse debate e seus rumos. Os resultados demonstraram que investigadores ao longo dos anos, desde que surgiram as primeiras bases de dados, continuam realizando estudos exploratórios para observar comportamentos de indexadores e de usuários, uso de ferramentas, processos, produtos e ambiente com diferentes contextos e tipologias textuais. Estudos de aplicação de novas metodologias, processos e produtos também são encontrados e demonstram novas possibilidades de soluções conciliadoras. Concluiu-se que a coexistência entre linguagem natural e linguagem controlada é necessária em sistemas de informação para a representação na indexação e na busca.
**Palavras-chave:** Palavra-chave. Linguagem natural. Linguagem controlada. Vocabulário controlado.

**ABSTRACT**
It is an essay to present the debate between natural language or controlled language in the contemporary perspective of the internet and web 2.0 in which the current databases, many of which are open access, such as repositories, are heterogeneous as to the organization and representation of information. The objective is to investigate, through literature review, the use of natural language or language controlled by information systems, in general, and by repositories, in a more specific way, to observe, evaluate and discuss the results, recommendations and reflections. In the qualitative analysis, 90 works were selected and, after previous reading, 73 scientific productions were chosen. Then, the reading was extensive to collect the objectives, methodology, results, conclusions and main points of view of the authors. A representative sample of 38 texts is distributed in 12 categories of analysis to facilitate the macro and micro understanding of this debate and its directions. The results showed that researchers over the years, since the first databases appeared, continue to carry out exploratory studies to observe behaviors of indexers and users, use of tools, processes, products and environment with different contexts and textual typologies. Studies on the application of new methodologies, processes and products are also found and demonstrate new possibilities for reconciling solutions. It was concluded that the coexistence between natural language and controlled language is necessary in information systems for representation in indexing and searching.
**Keywords**: Keyword. Natural language. Controlled language. Controlled vocabular.

*Artigo submetido em 10/10/2020 e aceito para publicação em 12/11/2020*

1 Parte de Projeto de Pesquisa com bolsa de Produtividade em Pesquisa do CNPq

2 Docente permanente no Programa de Pós-Graduação em Ciência da Informsação. Universidade Estadual Paulista Julio de Mesquita Filho, Brasil. Bolsista de Produtividade do CNPq. ORCID https://orcid.org/0000-0002-8239-7114. E-mail: fujita@marilia.unesp.br

## 1 INTRODUÇÃO

Existe um grande paradoxo de incertezas e diferenças não resolvidas na organização e representação da informação em bases de dados e catálogos de bibliotecas. De um lado, está o catálogo online cujas representações de seus recursos informacionais em metadados foram, ao longo de muitos anos, aprimorados por normalizações e instrumentos de controle de vocabulário que padronizaram e uniformizaram não só a representação da forma, como também do conteúdo dos recursos informacionais. Do outro lado, estão os repositórios, capazes de armazenar grandes quantidades de recursos digitais acompanhados de suas representações em metadados. Isso significa um avanço importante que proporciona a busca e o acesso ao documento sem que seja necessário verificar onde encontrá-lo. Essa adição da localização do conteúdo digital ao metadados tornou-se imprescindível aos usuários. Mas, o catálogo online, apesar de seu aprimoramento não consegue trazer junto aos metadados todos os arquivos de recursos digitais. Existem, assim, repositórios administrados por bibliotecas que não se articulam aos recursos do catálogo e vice-versa. Podemos pensar que é uma questão tecnológica por conta do uso de diferentes softwares. Mas, vejam que seria uma grande vantagem para a gestão documental ter todo o aprimoramento do catálogo, inclusive com a consistência de seus metadados, aliados ao grande potencial de armazenagem dos repositórios.

Entretanto, é na diferença que existe entre o uso de palavras-chave da linguagem natural em repositórios e o uso de vocabulários controlados no catálogo que se encontra o principal tema de debates da área de Organização e Representação da Informação. Desde o surgimento da primeira base de dados referencial de artigos de periódicos científicos até os dias atuais, a evolução tecnológica avançou de modo a garantir a comunicação científica porque todos sabemos que tão somente a publicação de artigos em um periódico não garante sua ampla comunicação. Assim, com o aumento da quantidade de periódicos científicos as bases de dados passaram a reunir todos os artigos publicados para que fosse possível a comunicação científica de tudo o que foi publicado em diferente periódicos.

Entretanto, o aumento exponencial de artigos impulsionou a representação consistente e pertinente do conteúdo dos artigos para que a recuperação fosse a mais adequada às necessidades de busca de seus usuários. A indexação intelectual foi utilizada e seus aprimoramentos, quanto à exaustividade e especificidade, foram intensamente avaliados para o alcance da precisão ou revocação na recuperação.

O aumento da quantidade de publicações também estimulou a indexação automática da linguagem natural de palavras do título e até do resumo com as propostas dos índices KWIC, KWOC e KWAC. Os índices rotados, como eram chamados, tinham o objetivo de criar índices com as palavras do título para oferecer pontos de acesso por assuntos aos usuários. A indexação semi-automática, parte automática e outra intelectual, foi também idealizada pela proposta do sistema PRECIS, no intuito de contextualizar a organização das palavras extraídas do conteúdo textual e não do título.

A construção desses índices passou a ser desnecessária com as tecnologias implantadas em documentos digitais e a maioria das bases de dados não precisaram pensar na indexação de cada artigo. Existem ainda bases de dados especializadas, construídas continuamente por redes de bibliotecas e outras instituições, que possuem política para realizar indexação com vocabulário controlado próprio em conjuntos selecionados de documentos (artigos científicos, trabalhos apresentados em eventos, dissertações e teses) como por exemplo, LILACS, MEDLINE, INIS ATOMINDEX. Isso significa que a mudança para a adoção de linguagem natural não foi total e que os critérios de seletividade bem como de rigor na indexação são obedecidos. Por outro lado, com o movimento do acesso aberto, surgiram bases de dados institucionais tais como repositórios e bibliotecas digitais que possuem novas configurações de organização e representação da informação condicionadas pelos softwares livres que afetam a forma de busca para recuperação. Isso é fato.

O percurso histórico brevemente exposto demonstra que sempre houve preocupação quanto à necessidade de realizar a indexação com controle de vocabulário, mas por outro lado, a grande quantidade publicações científicas produzidas sempre foi um problema de difícil resolução pela própria velocidade de publicação e disseminação. Assim, a linguagem natural, com todas as características presentes que impedem a precisão e revocação na recuperação, sempre foi o recurso natural economicamente disponível e, ao que tudo indica, continua sendo, assim como a indexação e o uso de vocabulários controlados por bases de dados politicamente situadas. Neste cenário, a palavra-chave é, sem dúvida, protagonista desse debate assim como as palavras do título de uma publicação.

A ideia presente nessa evolução é o debate entre linguagem natural ou linguagem controlada. Mas, quais foram os problemas e as soluções apresentados em cada caso? Quais as razões em se adotar uma ou outra. Quais as tendências retrospectivas e atuais? O objetivo desse trabalho de investigação ensaístico é investigar, por meio de revisão de literatura, o uso de linguagem natural ou de linguagem controlada por sistemas de informação, de modo geral, e por repositórios, de modo mais específico, para observar, avaliar e discutir os resultados, recomendações e reflexões. A ideia

deste ensaio é, também, discutir a coexistência de palavras-chave e descritores ou termos autorizados em sistemas de informação para representação de seus diferentes tipos de documentos e recursos de informação em diferentes áreas de conhecimento. Nesse sentido, o estudo não se propõe a definir uma posição unilateral entre a linguagem natural ou o controle de vocabulário ou realizar estudo comparado de vantagens e desvantagens entre as duas linguagens conforme Lopes (2002), mas apontar recomendações e tendências da revisão de literatura sobre o uso de cada uma ou de ambas para a representação na indexação e na busca.

## 2 LINGUAGEM NATURAL OU LINGUAGEM CONTROLADA: ASPECTOS HISTÓRICOS EVOLUTIVOS

Linguagem natural ou controlada é um tema de longa e permanente discussão na área de Organização do Conhecimento como demonstrado por Lopes (2002) em revisão de literatura cujos estudos citados abordam o uso das linguagens controlada e natural nas estratégias de busca e discute vantagens e desvantagens.

O estudo de Henzler, publicado em 1978, sobre comparação quantitativa de texto livre e vocabulário controlado utilizado na indexação e recuperação, realizou avaliação de consistência de indexação com 683 artigos da área biomédica. Os resultados obtidos mostram a necessidade de combinação ideal entre ambos.

Antes do estudo de Henzler, Foskett (1973) cita pesquisas em recuperação da informação realizadas na década de 50 com sistemas de informação especializados que adotaram listas de cabeçalhos de assuntos na indexação de documentos para obterem controle de vocabulário. O mais célebre foi o Projeto Cranfield realizado em 1957 pela Aslib que adotou metodologia de avaliação comparativa de quatro linguagens controladas, entre elas o Unitermo criado por Mortimer Taube. Os resultados comparados obtidos da indexação de documentos por diferentes indexadores que utilizaram cada um uma determinada linguagem, revelaram que todas tiveram a mesma eficiência e que os resultados do Unitermo foram ligeiramente melhores que as demais.

Segundo Lancaster (2002, p.50) o sistema Unitermo apresenta a característica de "[...] representar assuntos por palavras simples extraídas do texto dos documentos, sem nenhum tipo de controle". Chu (2003, p.6) considera que

> Unitermos podem, em certo sentido, ser considerados como palavras-chave de hoje porque ambos são derivados de documentos originais e nenhum esforço é feito para controle de

> vocabulário (por exemplo, verificar sinônimos e homógrafos). Normalmente, vários Unitermos são usados para representar um único documento, como na indexação de palavras-chave.

Em realidade o Unitermo foi criado para viabilizar um processo de indexação denominado *coordinate indexing* cujo objetivo era a recuperação da informação de documentos técnicos científicos a partir da coordenação de dois ou mais unitermos, o que comprova, em certa medida, a influência exercida pela recuperação na indexação e vice-versa.

Segundo Gomes (S.d), "Havia um pressuposto que as ideias seriam representadas por uma única palavra", o unitermo. A simplicidade do Unitermo provocava inúmeros problemas na época, sem os computadores e os softwares de busca, o que levou ao aparecimento do primeiro tesauro, elaborado pela *Armed Services Technical Information Agency* (ASTIA) em 1960 que, basicamente, tinha a proposta de estabelecer um controle da grande quantidade de termos simples extraídos dos documentos (LANCASTER, 2002).

Na área de arquivos, Paes (2004) cita o unitermo como um método de arquivamento para arquivos especiais e especializados onde após a análise dos documentos devem ser destacados os elementos identificadores que serão "[...] transcritos em fichas índices sob a forma de palavras-chave, quando os termos forem extraídos dos documentos analisados, ou descritores, quando utilizadas palavras constantes de um ou mais Thesauri técnicos (vocabulário controlado)" (PAES, 2004, p.90). Aquino e Aquino (2013) observaram em análise de 38 periódicos da área de ciências agrárias, publicados no período 1999-2011 que existem diferentes formas da escrita da seção "Palavras-chave" em artigos científicos tais como "Descritores", "Termos de indexação" e "Unitermos", o que demonstra a repercussão do Unitermo até os dias atuais.

Barité (2014, p.99) refere-se "[...] à difícil coexistência entre as técnicas da linguagem natural e as técnicas da linguagem controlada" e relata com base em Chu (2003) quatro fases na história da representação e da recuperação da informação que também nos ajudam a entender o debate. As fases tiveram início na primeira metade do século vinte até o surgimento da web 2.0 de 2000 em diante. O uso da linguagem natural era preferido na primeira metade do século vinte tendo em vista a simplicidade das primeiras listas de cabeçalhos de assunto existentes, mas os problemas de recuperação da informação provocados por termos sinônimos e homógrafos impediam o uso exclusivo da linguagem natural. Os vocabulários controlados na segunda metade do século vinte incorporaram procedimentos de controle que marcaram o desenvolvimento de outras listas de cabeçalhos de assunto, sistemas de classificação e tesauros com uso de métodos de construção. Com isso, o debate tornou-se mais

intenso entre os partidários da linguagem natural e do controle de vocabulário. Entretanto, de 1990 em diante surgiram novas técnicas de informática para recuperação da informação a partir da linguagem natural que conseguiam explorar o conteúdo dos documentos de texto completo sem controle de vocabulário. Com o advento da internet surgiram ferramentas de recuperação a partir da linguagem natural. A revolução da web 2.0, de 2000 em diante, está no uso social e colaborativo dos sistemas de informação que desenvolveram interfaces amigáveis ao usuário para representação e recuperação da informação com uso de linguagem natural. Os usuários de sistemas de informação atuais não precisam mais de intermediários para realizarem indexação e recuperação da informação como no início do aparecimento das primeiras bases de dados cujas interfaces de busca eram inexistentes e necessitavam de intermediação (LOPES, 2002).

As duas últimas fases, conforme Chu (2003), são ~~as~~ eras em rede na internet com avanços significativos para a redefinição da representação e recuperação da informação. Segundo Chu (2003, p.4), "Nunca antes na história da recuperação da Informação tantos usuários realizaram pesquisas online sem a ajuda de intermediários. " A esse respeito, Barité (2014) revela que a situação do controle de vocabulário em ambiente digital é regida por usuários autossuficientes que utilizam preferencialmente a linguagem natural, enquanto os responsáveis por portais e sites formalizam seus dados com critérios tradicionais de padronização.

Lu, et al. (2019) e Miguéis, et al. (2013), autores citados nesta revisão de literatura, concordam que, palavras-chave selecionadas por autores são multifuncionais porque tem sido utilizada para indexação, desenvolvimento de tesauros, marcação social, extração de palavras-chave, recuperação da informação, estudos bibliométricos e organização do conhecimento em diversos estudos. Além disso, os usuários esperam realizar buscas como se estivessem no Google e utilizam suas palavras-chave com a confiança de recuperarem conteúdos. Independentemente dessa alegada multifuncionalidade a linguagem natural tem difícil coexistência com a linguagem controlada (BARITÉ, 2014).

## 3 MÉTODO

A revisão de literatura foi realizada a partir de levantamento bibliográfico selecionado em bases de dados (BRAPCI, BDTI, Repositório Cruesp, SCOPUS, WoS, LISA, LISTA), metabuscador Google Scholar e Bibliografia da ISKO *Literature* sem delimitação temporal com uso dos termos "control vocabulary", "controle de vocabulário", "keyword", "palavra-chave", "controlled vocabulary" e "vocabulário controlado".

A seleção de referências para a revisão de literatura teve como critério a leitura do título e do resumo de cada texto para verificar o uso de linguagem natural ou de linguagem controlada e a coexistência de palavras-chave e descritores em sistemas de informação para representação de seus documentos.

Para a análise qualitativa da literatura, realizamos leitura prévia de 90 trabalhos e selecionamos 73 produções científicas compostas por teses, dissertações, artigos de periódicos, trabalhos completos em publicações de eventos e relatórios técnicos e de pesquisa. Na sequência, a leitura foi extensiva para coleta dos objetivos, metodologia, resultados, conclusões e principais pontos de vista dos autores que foram utilizados para descrição da pesquisa e comentários da autoria deste artigo. Com isso, foi possível definir uma amostra representativa de 38 textos presentes no debate distribuídos em 12 categorias de análise para facilitar a compreensão macro e micro desse debate e seus rumos.

A escolha desses textos incluiu qualquer contexto ou ambiente de sistema de informação, desde que houvesse a temática linguagem natural ou linguagem controlada. Cada categoria possui uma definição descritiva do propósito temático seguida da descrição de trabalhos analisados e ao final comentários analítico-críticos quando necessário.

As categorias de análise tiveram inspiração na organização da revisão de literatura de Gross, Taylor e Joudrey (2014) que destacaram como principais temas de discussão encontradas na literatura: *"Prevalência de busca com palavras-chave", "Vocabulário controlado é necessário para pesquisa acadêmica", "Dificuldades dos usuários com a pesquisa de assunto", "Razões contrárias a confiar na busca por palavras-chave", "Vocabulário controlado em campos específicos de estudo", "Soluções oferecidas", "Necessidade de vocabulário controlado, mesmo com texto completo disponível".*

Alguns desses temas possuíam subtemas significativos que representaram possibilidades investigadas, como: no tema *"Prevalência de busca com palavras-chave"* houveram duas possibilidades "confiar no vocabulário controlado na busca de palavras-chave" e "vocabulário controlado abandonado"; no tema "*Razões contrárias a confiar na busca por palavras-chave"* estão vinculados os subtemas "custo movido para os usuários" e "vocabulário controlado necessário para recursos não textuais"; e no tema *"Soluções oferecidas"* estão "vocabulário controlado e busca de palavras-chave", "sistemas de marcação do usuário", "uso dos termos de pesquisa do usuário para vocabulários controlados", "ferramentas protótipos" e "adição de tabelas de conteúdo e resumos".

A análise da literatura não teve como objetivo comparar resultados com a revisão de literatura de Gross, Taylor e Joudrey (2014) apenas tomou como parâmetro os temas para criação de categorias de análise, portanto, algumas categorias ciadas empiricamente têm semelhança com os citados temas.

## 4 O DEBATE SOBRE PALAVRAS-CHAVE DA LINGUAGEM NATURAL E DESCRITORES DE LINGUAGENS CONTROLADAS: RESULTADOS DA REVISÃO DE LITERATURA

Para obter uma visão macro organizamos as descrições dos estudos analisados a partir de 12 categorias: Análise de palavras-chaves atribuídas pelos autores; Análise de palavras-chave atribuídas por autores comparadas com vocabulário controlado; Vocabulário controlado em campos específicos de estudo; A busca de palavras-chave depende do vocabulário controlado como parte do sistema; Razões contra confiar na pesquisa de palavras-chave; Uso combinado de vocabulário controlado com pesquisa de palavras-chave; Uso combinado das linguagens natural e controlada; Vocabulário controlado necessário para recursos não textuais; Estudos sobre o uso de sistemas de marcação do usuário para indexação social; Uso dos termos de busca do usuários para vocabulários controlados; Interoperabilidade entre vocabulários controlados; e, Necessidade de vocabulário controlado na indexação e recuperação. A visão micro é revelada a partir do conjunto de estudos analisados em cada categoria de modo a demonstrar objetivo, metodologia, resultados e reflexões dos autores.

Os autores são os indexadores em sistemas de autoarquivamento atualmente existentes em repositórios e plataformas de gestão de periódicos. A indexação é realizada pelos autores com palavras-chave representativas do conteúdo que produziram. **A análise de palavras-chave atribuídas por autores** é investigada em estudos de Medeiros (2010), White (2013) Santos e Neves (2019), Santos, F. (2017), Lu, et al. (2019) e Freitas (2019).

Em análise de metadados de assuntos e resumos atribuídos por autores de artigos publicados em periódicos de Ciência da Informação no período de 1998 a 2008, Medeiros (2010) observou casos de sinonímia e polissemia em resultados da análise do metadado de assunto que provocaram baixa revocação e baixa precisão na recuperação da informação e conclui que a padronização é necessária na representação e na recuperação. A análise do metadado assunto de monografias depositadas no repositório da UFRN foi objetivo da pesquisa desenvolvida por Santos, R. (2017) que observou os mesmos problemas relacionados a sinonímia e polissemia.

White (2013) parte do princípio de que "Os repositórios científicos criaram um novo ambiente para o estudo de questões tradicionais da Ciência da Informação" porque a interação entre os termos de indexação fornecidos pelos usuários e os vocabulários controlados fornece mais questões para a continuidade do debate e levanta uma grande área de estudo em que as propostas precisam ser inovadoras. Desse ponto de vista, foi realizada pesquisa que examinou as práticas de organização da

informação de profissionais da informação e cientistas ao aplicar metadados e termos de assuntos para descrever o conjunto de dados científicos, além de examinar o mapeamento de quatro vocabulários controlados gerais e específicos mediante busca por palavras-chave. Para isso, o estudo usou uma abordagem de métodos mistos de triangulação simultânea para estudar os metadados e a aplicação de termos de assuntos por cientistas e profissionais da informação a conjuntos de dados. Os resultados do mapeamento de vocabulário indicam que, com base nos quatro subtipos de termos de assuntos examinados por este estudo (espacial, temporal, tópico e científico), o LCSH teve a melhor cobertura de termos de assunto para termos tópicos. Para termos científicos, o ITIS representou nomes científicos muito bem e teve a maior pontuação média de todos os vocabulários. Os resultados obtidos sugerem que os repositórios e os profissionais da informação não devem abandonar o uso de vocabulários controlados.

A pesquisa de Santos e Neves (2019), realizada para analisar as práticas de indexação no Repositório Institucional da Universidade Federal do Rio Grande do Norte, avaliou descritores atribuídos por autores no metadado assunto para descrição do conteúdo das produções intelectuais e obteve como resultado a presença de emprego de frases, termos com polissemia, ocorrência de erros ortográficos, presença de descritores abrangentes, abreviações nos termos, uso de siglas, entre outros aspectos observados. Diante desses resultados, foi recomendado não excluir os termos provenientes da coleta automática de palavras-chave utilizadas na indexação e substitui-las por outra forma de representação, entretanto, indicaram que é necessário utilizar correção automática para analisar continuamente e adicionar remissivas com auxílio de vocabulários controlados. Concluiu-se que a política de indexação deve ser considerada para definir práticas de representação da informação.

Freitas (2019) aplicou metodologia de avaliação intrínseca qualitativa da indexação de teses e dissertações do Repositório Institucional da UFSCar por autores/usuários e bibliotecários cujos resultados revelaram "concordância entre as indexações avaliadas, sendo a indexação do usuário mais exaustiva, porém sem a mesma precisão da indexação realizada pelo bibliotecário especialista." A autora recomenda como alternativas para a falta de precisão dos termos a disponibilização de vocabulários controlados aos autores para a atribuição de termos durante o autoarquivamento de publicações, teses e dissertações, a validação do metadados de assunto por bibliotecários e a elaboração de política de indexação para a representação e recuperação por assuntos em repositórios.

Santos, F. (2017) investigou comparativamente a coerência semântica na indexação de 10 artigos de periódicos da área de Saúde Pública publicados entre 2012-2014 e indexados no Repositório de

Produção Científica da Escola Nacional de Saúde Pública Sergio Arouca (ENSP) da Fundação Oswaldo Cruz (Fiocruz) com uso de técnicas e ferramentas tecnológicas bibliométricas. Os resultados obtidos revelaram baixo grau de coerência semântica na maioria dos artigos estudados.

Lu, et al. (2019) consideram, porém, que poucos estudos investigaram os padrões de palavras-chave selecionadas pelo autor em artigos científicos. O método de seleção de palavras chave utilizado pelo autor do artigo é o estudo que os autores desenvolvem com base na função semântica da palavra-chave em artigos científicos. A investigação é realizada a partir das funções dos termos (TF) de palavras-chave em artigos do *Journal of Informetrics* que incluem, "tema de pesquisa", "método de pesquisa", "objeto de pesquisa", "área de pesquisa", "dados" e "outros". Os dados foram processados manualmente com base em análise de conteúdo e utilizaram metodologia inovadora. As principais conclusões obtidas indicam que há uma efetiva relação entre a classificação de uma palavra-chave e sua função de termo e que as palavras-chave "temas de pesquisa" e "métodos de pesquisa" são as mais frequentes.

O aprimoramento de vocabulários controlados é uma meta constante que depende do desenvolvimento de estudos sobre **Análise de palavras-chave atribuídas por autores comparadas com vocabulário controlado** como investigada por Migueis e Neves (2013), Miguéis, et al. (2013) e Smiraglia (2013, 2015).

Palavras-chave atribuídas por autores em 207 artigos da área de Ciências da Saúde depositados no repositório da Universidade de Coimbra e indexados na base de dados MEDLINE foram comparadas com termos da *Medical Subject Headings* (MeSH) no estudo de Miguéis e Neves (2013). Os resultados obtidos demonstram que apenas cerca de metade das palavras-chave do repositório estão incluídas nas versões publicadas pelas editoras e em quantidade inferior aos termos empregados pela MEDLINE. Do total de palavras-chave identificadas dois terços coincidem ou tem relação associativa com os descritores do MeSH.

Como editor do periódico *Knowledge Organization* (KO), Smiraglia (2013, 2015) descreve em seus dois editoriais a experiência realizada com a indexação dos artigos publicados em dois fascículos do volume 40 de 2013 e os primeiros artigos de cada fascículo do volume 41 do ano de 2014 com as palavras-chave reais extraídas dos textos para a comparação com palavras-chave retiradas da Web of Science (WoS) e da Biblioteca e Resumos de Ciência e Tecnologia da Informação da EBSCOHost com Texto Completo (LISTA). A indexação nos dois estudos de caso foi realizada pela editoria do periódico e não pelos autores dos artigos porque palavras-chave não são solicitadas quando da

submissão e preenchimento de metadados. Nos dois casos a seleção de palavras-chave foi realizada com base no conteúdo real dos artigos. Pela análise dos quadros comparativos observa-se que não há correspondência de termos idênticos entre as palavras-chave fornecidas pela KO e as palavras-chave dos dois indexadores. Sobre isso, Smiraglia (2015) relata que quando autores enviavam seus textos atribuíam palavras-chave, sem que fosse solicitado, mas que observava não ter correspondência com o conteúdo do artigo. Sob o prisma do processo de indexação, propriamente dito, palavras-chave devem ser representativas do conteúdo analisado e não somente atribuídas por uma questão de importância ou de melhor divulgação em bases de dados, conforme ponderou Smiraglia (2013, p.155) que decidiu continuar a prática editorial usando palavras-chave representativa do conteúdo dos artigos porque entende "[...] que a única coisa que melhora a recuperação é indexação formal."

Por outro lado, Miguéis et al. (2013) realizaram análise comparada de termos da Medical Subject Headings com palavras-chave atribuídas por autores da Universidade de Coimbra em 182 artigos publicados em periódicos internacionais entre 1996 a 2012 da área das Ciências Farmacêuticas depositados até final de 2012 no Repositório da Universidade. Os resultados obtidos revelaram que a quantidade de palavras-chave atribuídas por artigo é menor do que os termos do MeSH. Entretanto, metade do total de palavras-chave apresentaram relações de equivalência ou associativas em comparação com os termos do MeSH o que levaram os investigadores a concluírem que isso representa uma influência direta ou indireta na escolha dos descritores e que os autores/indexadores são intervenientes ativos no processo de representação e recuperação da informação através de suas palavras-chave e promotores diretos da visibilidade de seus trabalhos que reverterá em aumento de citações e impacto da produção científica da universidade.

O principal objetivo de bases de dados, além de armazenar, é propiciar condições favoráveis ao acesso dos conteúdos e recursos informacionais disponíveis. Em bases de dados com grandes quantidades e diversidade de conteúdos torna-se estratégica a busca por assunto que ofereça ferramentas semânticas para representar necessidades de informação de seus usuários. Vocabulário controlado é necessário para pesquisa por assunto porque a **busca de palavras-chave depende do vocabulário controlado como parte do sistema** em estudos desenvolvido por Hanrath e Radio (2017) e Gollub (2016, 2018). Entretanto, o investimento em vocabulários controlados como parte do sistema é custoso em termos de estrutura de recursos humanos e tecnológica e a solução pode ser propiciar o acesso ao vocabulário controlado nas buscas por assunto. Além disso, os resultados obtidos (GOLLUB, 2016, 2018) demonstram que há restrições quanto ao uso do vocabulário controlado e que é preciso aprimoramento com atualização tecnológica.

Hanrath e Radio (2017) estudaram o comportamento de busca dos usuários em repositórios institucionais, comparando os tópicos das consultas levantadas e os metadados dos títulos e assuntos, com foco em um específico. Eles consideram que grande parte das consultas levantadas nos repositórios são de natureza temática e que lidar com a descrição temática com mais cuidado pode melhorar as pesquisas. O estudo apresenta método de análise do comportamento de pesquisa do usuário para ajudar os administradores de repositório a determinar se devem investir na aplicação de vocabulários controlados por assunto para recuperar informações em repositórios. A taxa comparativamente alta em que as consultas de pesquisa do usuário correspondem a termos de assunto existentes sugere novamente que a atenção à descrição do assunto pode melhorar a visibilidade do conteúdo nos repositórios. Aplicar um vocabulário controlado ao conteúdo dos repositórios pode ser desafiador. Mas a aplicação retroativa de um vocabulário controlado, bem como o projeto de mecanismos para sua implementação contínua, pode consumir tempo e recursos consideráveis. O benefício de aplicar vocabulário controlado aos termos do assunto deve ser pesado em relação aos seus custos, incluindo o custo de oportunidade de tempo e recursos que poderiam ser usados para melhorar outros metadados nos repositórios.

Gollub (2016) relata estudo exploratório sobre acesso por assuntos em catálogos de bibliotecas da Suécia cujos resultados indicam que o acesso ao assunto não é abordado sistematicamente, que é limitada a aplicação de sistemas de organização do conhecimento em coleções digitais e que não existe a operação de mapeamento entre os diferentes sistemas de organização do conhecimento utilizados nos diferentes bancos de dados integrados de bibliotecas e serviços de informação comerciais. A autora sugere o uso de marcação social e indexação automática de assuntos a serem implementados em continuidade ao estudo anterior de Golub, Lykke e Tudhope (2014).

Na mesma linha das pesquisas anteriores, modelos conceituais contemporâneos e código de catalogação, Golub (2018) analisou a busca por assunto de três serviços de descoberta mais comuns usados em bibliotecas utilizando a proposta de análise de 18 características de interfaces contemporâneas. Os resultados indicam que, apesar dos catálogos terem indexado suas coleções individuais com auxílio de vocabulários controlados e de mapeamento de número significativo de vocabulários controlados, o acesso ao assunto em interfaces contemporâneas é menos do que o ideal. A autora recomenda a construção de padrões e diretrizes baseados em pesquisas.

A pesquisa por palavras-chaves, além de basear-se na linguagem natural, é uma estratégia certamente muito natural hoje em dia, principalmente quando temos uma ferramenta como o Google para nos ajudar a encontrar qualquer informação ou conteúdo sem que seja necessário preocupar-se

com a forma e conteúdo do que digitamos na estratégia de busca. Contudo, **existem razões contra confiar na pesquisa de palavras-chave** apresentada no estudo realizado por Materska (2016).

Com o objetivo de descrever o estado atual dos repositórios de universidades polonesas no contexto de ferramentas terminológicas a pesquisa desenvolvida por Materska (2016) com grupo de 12 repositórios acadêmicos e 3 departamentos, centralizados em diretório nacional como único ponto de acesso, analisou as atividades baseadas em palavras-chave de dois grupos de usuários do repositório, dos pesquisadores durante o autoarquivamento para descrever suas publicações e as dos usuários finais ao pesquisar, navegar e recuperar o conteúdo do repositório. Materska (2016) esclarece que o foco principal da investigação está em palavras-chave, termos de assunto selecionado para descrever a tematicidade dos objetos científicos no repositório. Nessa perspectiva, os resultados da pesquisa levam a autora a considerar que a falta de controle de autoridade das palavras-chave adicionadas pelos pesquisadores provoca a presença de vários erros muito óbvios como formas do mesmo termo com erros de grafia ou no singular e no plural colaborando para que haja ampla dispersão do vocabulário na representação quando cada termo é usado para indexar 1 ou 2 títulos apenas. Na comparação entre pesquisadores e usuários os repositórios universitários são explorados, em geral, com mais eficiência pelos pesquisadores que conhecem os nomes dos autores e títulos dos artigos. A autora conclui que é preciso melhorar as descrições de conteúdo e funções de pesquisa de repositórios de universidades polonesas. Por outro lado, a autora chama a atenção para o fato de que a pesquisa sobre repositórios sob a perspectiva terminológica é modesta e que essa situação não surpreende porque os desenvolvimentos dos repositórios tiveram foco inicial nas dimensões organizacional, legal, promocional e outas dimensões práticas e não nos aspectos terminológicos, navegação, pesquisa e descoberta de conteúdo.

O uso de **Vocabulário controlado em campos específicos de estudo** apresenta diferenças que precisam ser resolvidas e, ao mesmo tempo, seriam a solução para a combinação entre linguagem natural e linguagem controlada. Nesse sentido, as pesquisas de Maurer e Shakeri (2016), realizada com as áreas de Artes e Humanidades, Ciências Sociais, Ciências, Tecnologia, Engenharia e Matemárica, de Gollub, et al. (2020), realizada com a área de Ciências Humanas, e de Han, et al. (2016) que demonstra a necessidade de uso de vocabulários específicos, são importantes para se obter bons resultados na representação para indexação e recuperação intermediadas com uso da linguagem natural.

Os resultados de pesquisa (MAURER e SHAKERI, 2016) realizada no catálogo da Biblioteca da Kent State University revelam que, em média, mais palavras-chave atribuídas pelo autor e mais

cabeçalhos de assunto da Biblioteca do Congresso atribuídos pelo catalogador foram atribuídos a trabalhos emergentes das artes e humanidades do que a trabalhos emergentes das ciências sociais e ciências, tecnologia, engenharia e matemática (CTEM). As disciplinas CTEM receberam quantidade menor de metadados tópicos, porque menos cabeçalhos de assuntos de nomes/títulos, geográficos e corporativos foram atribuídos. Concluem que a literatura demonstra escassez de informações a respeito da relação entre o número de palavras-chave fornecidas pelo aluno autor e qualquer experiência na submissão de artigos.

A indexação de assuntos em humanidades é investigada por Gollub, et al. (2020) com o objetivo de apresentar a situação atual do uso dos termos de indexação de assuntos em artigos de periódicos de Ciências Humanas com referência às necessidades de acesso a assuntos de pesquisadores de Ciências Humanas. A comparação de metadados de assuntos foi realizada com amostra composta por 649 artigos de periódicos de Humanidades extraídos do repositório de uma universidade pública da Suécia, dotada de políticas locais e nacionais, dos quais 321 foram também localizados na Scopus, banco de dados internacional de resumos e citações com política maior e mais abrangente. Os resultados obtidos pelo estudo mostram que os objetivos bibliográficos estabelecidos pelos dois bancos bibliográficos quanto ao acesso por assunto aos artigos de periódicos de humanidades não são adequadamente garantidos. Conforme conclusões há evidências de que nenhum vocabulário controlado de Humanidades é usado e os artigos que usam termos de indexação são de vocabulários controlados de outras áreas de conhecimento tais como EMTREE, MeSH e GEOBASE; as categorias do repositório se destinam principalmente à análise estatística; não há mapeamento entre os vocabulários e isso produz duplicatas e impede o uso de termos de indexação de um mesmo vocabulário em todos os recursos de informação; os autores não tem qualquer treinamento ou orientação sobre indexação e ignoram a influência na recuperação em veículos de comunicação científica. As recomendações do estudo referem-se à marcação social com sugestões produzidas automaticamente, de preferência derivadas de vocabulário controlado para aprimoramento do campo de assunto, conforme resultados positivos obtidos em estudos que investigaram o aprimoramento da marcação social com sugestões da Classificação Decimal de Dewey (KHOO, et al., 2012; GOLUB, LYKKE, TUDHOPE, 2014).

Han, et al. (2016) consideram que nas últimas décadas as bibliotecas experimentaram uma evolução em seus serviços de recuperação; fizeram a transição de OPACs para serviços de descoberta em escala da web que permitem o acesso a ambos os recursos no OPAC, bem como a artigos e capítulos disponíveis a partir de assinaturas de banco de dados núcleo, cujos recursos são descritos

com termos de assunto mais específicos do que aqueles oferecido no LCSH (Larson 1991); segundo, as bibliotecas agora estão lidando com mais e mais metadados criados por não catalogadores (por exemplo, metadados fornecidos pelo autor), muitas vezes usando termos de assunto não disponíveis no LCSH ou em outros vocabulários controlados estabelecidos; e em terceiro lugar, as bibliotecas e fornecedores ainda não desenvolveram boas práticas para fornecer aos usuários serviços de acesso a tópicos específicos para cada disciplina. Esses autores consideram que as palavras-chave podem se alinhar melhor com vocabulários controlados específicos de uma disciplina.

Ao longo do tempo propostas inovadoras surgiram para aproveitamento tanto da linguagem natural quanto da linguagem controlada considerando que ambas oferecem vantagens. Uma das principais **soluções oferecidas** referem-se ao **uso combinado de vocabulário controlado com pesquisa de palavras-chave** pelos estudos de Borst (2012), Hartley e Kostoff (2003), Silva e Lima (2015), Tartarotti (2019), Gross, Taylor e Joudrey (2015).

Borst (2012) examinou os arquivos de log do repositório da Biblioteca Nacional Alemã de Economia para verificar o uso efetivo e o impacto dos termos controlados durante a recuperação por usuários tendo em vista que o comportamento de pesquisa seria potencialmente influenciado pela sugestão automática ou pela expansão de termos derivados da literatura. O repositório indexa publicações científicas de acesso aberto com uso do Standard Thesaurus for Economics (STW) e, ao mesmo tempo, é integrado ao repositório de assuntos da instituição para sugerir automaticamente palavras-chave durante a indexação e recuperação e para expandir automaticamente as consultas de pesquisa sob demanda. Os resultados obtidos demonstraram que cerca de um terço de todas as consultas ao repositório contém termos de pesquisa de vocabulários controlados com uma pequena maioria de termos do STW e que houve "rolagem" das páginas de referências encontradas na busca por documentos relevantes o que sugere confiança nos resultados. Borst (2012) sugere técnicas linguísticas para o mapeamento de termos não controlados como, por exemplo, a lematização além de ações para melhorar a recuperação com vocabulário controlado que podem afetar a interface do usuário:

> A expansão do termo de pesquisa deve ser realizada de forma discreta, mas ainda visível e transparente. Os termos de pesquisa sugeridos podem ter o título correspondente ("Termos relacionados à sua consulta e associados a documentos") e renderizados de acordo com a frequência.
>
> A rolagem excessiva dos conjuntos de resultados geralmente deve ser evitada e resolvida com a introdução de classificação de colunas, pesquisa em cascata e filtros.

> Como um número significativo de termos não controlados pertence a outras – categorias como nomes, números e títulos de documentos, isso deve ser melhor suportado pelo sistema de informações que responde. Para esse fim, sugerimos uma infraestrutura de dados baseada em dados de autoridade dessas categorias. Mais concretamente, uma entrada não controlada seria comparada simultaneamente com diferentes conjuntos de dados de autoridade, sugerindo que o usuário refinasse sua pesquisa na forma de 'Você quis dizer a pessoa / conceito / obra / título?' Depois de escolher uma categoria, um campo interno de pesquisa seria acionado.

Hartley e Kostoff (2003) verificam em sua pesquisa a indagação "quão útil são palavras-chave em periódicos científicos? " e partem do princípio que os conteúdos dos artigos são representados pelo título, resumo e palavras-chave em bancos de dados no auxílio à recuperação da informação e na web por autores, leitores, indexadores e sistemas de informação em geral. O conjunto de palavras-chave indicam os principais conceitos e campos de assuntos do artigo e, para isso, consideram que uma lista abrangente de palavras-chave poderia ser disponibilizada tal como a taxonomia do Medical Subject Headings (MeSH) utilizada pelo MEDLINE. Sugerem que essa taxonomia poderia incluir o número de registros que cada termo representa no MEDLINE para que o autor tenha o parâmetro de garantia de uso.

O conceito de navegação facetada na pesquisa realizada por Silva e Lima (2015), mais atualizada que a pesquisa de Hartley e Kostoff (2003) está vinculada às taxonomias cujas categorias de assuntos são expostas em lista alfabética abaixo da interface de busca como mais uma opção de navegação que pode ser combinada com outras categorias e palavras-chave. Essa lista alfabética de categorias de assunto, denominadas facetas, são extraídas de uma taxonomia de termos autorizados e controlados. Na pesquisa de Silva e Lima (2015) foi proposta uma interface de busca que combina a navegação facetada e a busca por palavras-chave em um catálogo web facetado de empresas. A avaliação foi realizada com teste de usabilidade para tarefas de recuperação da informação de forma livre para que o usuário decidisse a forma de busca, palavras-chave, ou navegação faceta ou a combinação de ambas. A preferência do usuário foi pela busca com palavras-chave (50%), a navegação facetada foi utilizada em um terço das buscas (29%) e a combinação de ambas em apenas 7%. Os autores concluem que é preciso fazer outras pesquisas para ajustes na interface de busca para a combinação de vocabulário controlado e palavras-chave da linguagem natural.

Tartarotti (2019) realizou estudo de avaliação da indexação e da recuperação documental por assuntos de forma comparada entre linguagem natural e linguagem controlada no Repositório Institucional da UNICAMP. A avaliação da indexação foi realizada com metodologia de interconsistência *inter-autor-bibliotecário* para a comparação dos índices de consistência entre os assuntos atribuídos

pelos autores (palavras-chave) e pelos catalogadores-indexadores (descritores). A média de interconsistência entre autores e catalogadores-indexadores foi baixa em todas as quatro grandes áreas do conhecimento tendo em vista que a maior média alcançada foi de 10,02% na área de Ciências Exatas. Os resultados da avaliação comparada da recuperação por assuntos em linguagem natural e linguagem controlada, revelam baixo índice de precisão tanto em linguagem natural (12,97%) quanto em linguagem controlada (9,93%) e entre as duas opções a linguagem natural tem maior índice de precisão. Tais resultados levam à várias conclusões, mas uma delas é, sem dúvida, o aprimoramento e atualização do vocabulário utilizado. Por isso, a autora indica recomendações no sentido de criar uma política de indexação institucional e construir um vocabulário controlado para a UNICAMP e sua disponibilização para uso combinado na indexação, por autores e catalogadores/indexadores, e na recuperação com adoção de melhorias na ferramenta de busca.

A investigação sobre a função do vocabulário controlado versus palavras-chave em catálogo online levada a cabo por Gross, Taylor e Joudrey (2015) é uma continuação da pesquisa de Gross e Taylor (2005) cuja conclusão é de que mais de um terço dos registros recuperados em buscas por palavras-chave seriam perdidos sem cabeçalhos de assunto. Após revisão de literatura sobre o debate vocabulário controlado versus palavras-chave o estudo em questão replicou o processo de pesquisa no mesmo catálogo online da University of Pittsburgh Libraries com adição de metadados enriquecidos automatizados, como índices e resumos. A conclusão do estudo de replicação com metadados enriquecidos após obtidos os resultados é de que o percentual médio de 27% dos acessos seria perdido na ausência de cabeçalhos de assunto e em buscas limitadas à língua inglesa seriam 24,8% que, comparado aos resultados de 35,9% de resultados perdidos do estudo anterior, representam 11,1% menos do que sem aprimoramento. A revisão de literatura realizada demonstrou ainda que a inclusão do vocabulário controlado em registros de metadados permite vantagens para controle de sinônimos, diferentes grafias, referências para temos obsoletos e diferentes significados de um mesmo termo, além de referências hierárquicas.

A combinação de palavras-chave e descritores é uma discussão presente na literatura e revela uma tendência cujas vantagens e desvantagens são influentes em decisões quanto à representação e recuperação da informação. Estudos de Hartley e Kostoff (2003), Kip, (2006), Bacha e Almeida (2013), Zavalina (2014), Sassen (2017) e Armani et al. (2000) realizaram análises comparadas de palavras chaves e descritores de assunto quanto à possibilidade de **Uso combinado das linguagens natural e controlada** como outra **solução oferecida**

Bacha e Almeida (2013) observaram em estudo sobre o uso combinado das linguagens natural e controlada, que práticas emergentes de melhoria na criação e descrição de metadados de assunto, usando mutuamente as duas linguagens, enriquece as pesquisas e facilita o acesso a objetos digitais. A proposta de Hartley e Kostoff (2003) de disponibilização da taxonomia do MeSH, citada anteriormente, é uma opção interessante pelo fato de mediar a combinação entre palavras-chave e descritores durante a submissão de artigos de periódicos por autores.

Com a proposta de examinar o contexto de indexação online, Kip (2006) analisou os resultados da indexação na perspectiva de três grupos diferentes de indexadores: de usuários, autores e intermediários. Para isso, foram coletadas e analisadas palavras-chaves de 165 artigos de periódicos marcados no "CiteULike" por meio de análise estatística e comparação de termos entre os três grupos para examinar semelhanças e diferenças entre os três conjuntos de tags. Diferenças foram observadas no contexto das palavras-chave dos três grupos e os resultados indicados demonstram que o maior número de rótulos fornecidos pelos usuários para um único artigo foi 21, pelos autores 10 e por intermediários 12. Mais de 60% dos artigos marcados tinham entre 1-3 tags, 4-6 palavras-chave do autor e 3-5 descritores intermediários atribuídos.

Os resultados de estudo (ZAVALINA, 2014) que trata da complementaridade entre metadados de assunto em texto livre e descritores a partir de um vocabulário controlado em três bibliotecas digitais de grande porte, demonstram empiricamente que registros de metadados mais detalhados ao nível da coleção, que incluem metadados de assunto de texto livre e vocabulário controlado, permitem uma representação mais completa do conteúdo intelectual dos objetos de informação e, finalmente, melhoram o acesso do assunto aos usuários.

Na perspectiva da prática profissional, uma análise dos registros de catálogo de bibliotecas acadêmicas da Association of Research Libraries em pesquisa realizada por Sassen (2017) para verificar como as bibliotecas fornecem acesso aos assuntos, bem como aos nomes dos departamentos acadêmicos e consultores, revelou que essas informações são registradas com mais frequência em notas e pontos de acesso não controlados do que em pontos de acesso autorizados para fornecer esse acesso. Conclui que a catalogação poderá ser completada mais rapidamente se notas ou pontos de acesso não controlados forem usados para registrar nomes e assuntos, contudo, essas práticas locais devem ser consideradas no contexto da recuperação da informação.

A linguagem natural e o **Vocabulário controlado necessário para recursos não textuais** são combinados durante a representação de mídias mistas em proposta de Armani, et al. (2000). Indexação

por similaridade é a proposta de Armani, et al (2000) para o acesso e recuperação de mídias mistas (texto, imagens, som, vídeo) com conteúdo representado por palavras-chave em bibliotecas multimídias temáticas mediante anotações semânticas semelhantes em metadados. As anotações seguem um processo de duas etapas: a anotação da primeira etapa é uma legenda feita com Linguagem Natural em forma de texto curto para descrever o conteúdo semântico do documento de mídia e na segunda etapa as legendas serão convertidas semiautomaticamente em anotações finais mediante representação com uso da linguagem NKRL *(Narrative Knowledge Representation Language)* que consiste de conjunto de conceitos e instâncias de conceitos (indivíduos) e conjunto de estruturas mais complexas (ocorrências). Os autores consideram que o processo de anotação em duas etapas garante flexibilidade à consulta com uso de palavras-chave e, ao mesmo tempo garante a representação uniforme por linguagem controlada.

A internet e, depois, a web 2.0 propiciaram a interação direta do usuário com os sistemas de buscas online na web. Surgiram propostas para que essa interação fosse socializada com os demais usuários no contexto de sistemas de informação. Dessa forma, o usuário pode atribuir termos descritivos à textos ou imagens e compartilhar com outros usuários. **Estudos sobre o uso de sistemas de marcação do usuário para indexação social** é investigada por Santos, R. (2016), Santos e Corrêa, (2015), Khoo, et al.(2012), Golub, Lykke, Tudhope (2014), Viana (2020) e Kipp, (2006) como outra **solução oferecida** como forma de aprimorar a representação na indexação e recuperação.

Em investigação sobre indexação social, Santos, R. (2016) avalia possível aplicação em base de dados referencial de artigos de periódicos em Ciência da Informação (BRAPCI) de três modelos colaborativos de indexação social identificados na revisão de literatura, modelo de Representação Iterativa, modelo colaborativo de indexação Facetlog e modelo colaborativo baseado em tags categorizadas com a finalidade de minimizar os problemas da indexação com linguagem natural de palavras-chave. Os resultados dessas avaliações demonstram que é possível a revisão dos termos dos metadados de assuntos para diminuir erros ortográficos e duplicações, além de minimizar a participação do indexador na atividade de categorização dos termos e definição de remissivas entre termos. Recomenda a elaboração de instrumento de vocabulário controlado com base nas garantias literária, estrutural e de uso para auxílio aos autores.

Em Santos e Corrêa (2015) os modelos colaborativos de indexação social para aplicabilidade em bibliotecas digitais são investigados por meio de revisão de literatura para identificar estudos que proponham modelos de integração da folksonomia em metadados. Com a descoberta do modelo

de representação iterativa de Santarém Segundo (2010) e do modelo colaborativo para indexação e busca de registros em catálogo web facetado (Facetlog) de Silva (2013) os autores observaram que os modelos controlam o nível de liberdade do usuário na atribuição de tags com maior grau de significado em relação ao objeto depositado.

Khoo et al. (2012) descrevem trabalho em andamento sobre proposta de desenvolvimento de ferramenta de Interface de Indexação de Documentos e Marcação Semântica para bibliotecas (DISTIL) para suporte de pesquisa federada de coleções cruzadas em Humanidades e Ciências Sociais tendo em vista a necessidade de infraestrutura unificada de informações para bibliotecas digitais. A ferramenta DISTIL propõe oferecer apoio à interoperabilidade a partir da geração de tags da Classificação Decimal de Dewey por meio de metadados individuais que poderiam ser usadas para a navegação entre coleções.

Na mesma direção de aplicação da Classificação Decimal de Dewey (CDD) como sistema de organização do conhecimento para aprimorar a marcação social e por consequência melhorar a indexação e recuperação por assuntos, Golub, Lykke, Tudhope (2014) realizaram pesquisa com estudantes de política para investigar o aprimoramento da marcação social com sugestões da CDD em teste de usuário de comparação da marcação social e marcação social enriquecida com sugestões da CDD. Para isso, cada estudante recebeu quatro tarefas, nas quais um total de 60 recursos informacionais foram marcados em duas configurações diferentes, uma com marcas sociais não controladas e outra com marcas sociais não controladas acompanhadas de sugestões do vocabulário controlado da CDD e mapeamentos dos cabeçalhos de assunto da Library of Congress Subject Headings. Os resultados obtidos revelaram a importância da marcação social acompanhada de sugestões de vocabulário controlado para indexação e recuperação no que se refere à indicação de ideias para escolha de tags e aumento de pontos de acesso.

Viana (2020) realizou pesquisa exploratória com repositórios da Rede de Repositórios de Dados Científicos do Estado de São Paulo e constatou a inexistência efetiva da prática de indexação social ou Folksonomia nos repositórios de dados embora disponham de infraestrutura tecnológica para o *tagging*.

A linguagem natural por usar termos livres apresenta maior diversidade e rápida atualização embora sem controle de vocabulário. O **Uso dos termos de busca do usuário para vocabulários controlados** é, sem dúvida, uma solução propícia ao aprimoramento de vocabulários controlados de áreas especializadas como demonstrado pela pesquisa de Souza (2010).

Souza (2010) relata experiência de adoção e controle de termos livres na indexação por palavras-chave pelo sistema da Agência de Informação Embrapa que desenvolveu um aplicativo integrado à ferramenta de catalogação do Sistema Gestor de Conteúdo para inserção de novos termos e de controle de termos autorizados pelo sistema no Banco de Termos Autorizados (BTA). Projetado para atender à necessidade de normalização e uso do vocabulário controlado, o BTA é uma ferramenta de uso do catalogador. O relato considera estratégica a integração de bases de autoridades com base de recursos informacionais porque elimina redundâncias e inconsistências na descrição do recurso e facilita a recuperação. Em desenvolvimento futuro pretendem o serviço web para o BTA tendo em vista a interação com outros servidores da web.

A proposta de mapeamento para **Interoperabilidade entre vocabulários controlados** especializados é imprescindível em áreas de pesquisa avançada que necessitam de controle de vocabulário continuamente atualizados. Rowel (2013) e Zhang, et al. (2015) desenvolveram pesquisas associadas à repositórios de dados que contemplam a integração de vários vocabulários controlados em ferramentas destinadas a selecionar termos de busca.

Conforme Rowel (2013), vocabulários controlados facilitam a interoperabilidade, mas apresentam desafios relacionados a custo, usabilidade e interdisciplinaridade. O projeto *Helping Interdisciplinary Vocabulary Engineering* (HIVE) visa enfrentar alguns desses desafios, fornecendo uma abordagem para a integração de vários vocabulários controlados. Em pesquisa para o desenvolvimento do HIVE, foi implementada pesquisa na Web direcionada a funções associadas a repositórios de dados - contribuidores de dados, curadores de dados, administradores de DataNet e desenvolvedores de repositórios - em relação ao uso de vocabulários controlados (ROWELL, 2013).

Rowell, em co-autoria de artigo com Zhang et al. (2015), realizaram pesquisa para examinar o uso controlado do vocabulário e os recursos de aplicativos específicos para dados científicos. Para os autores “Vocabulários controlados são sistemas semânticos úteis para organizar e acessar recursos - e para apoiar a interoperabilidade semântica entre descrições de objetos e repositórios” (ZHANG, et al. 2015, p.6). A ferramenta HIVE (*Helping Interdisciplinary Vocabulary Engineering*), elaborada por Rowell (2013), viabiliza o acesso a vários vocabulários controlados ao mesmo tempo e de forma dinâmica. Os resultados demonstram que os participantes da pesquisa consideram os vocabulários controlados como ferramenta valiosas e úteis que proporciona acesso a vários dados. Entretanto, entre os resultados, descobriu-se que há falta de ferramentas avançadas nos repositórios de dados para selecionar termos de assuntos de vários vocabulários controlados, anotar termos de vocabulários controlados e gerar termos automaticamente

A **Necessidade de vocabulário controlado na indexação e recuperação** é discutida em trabalhos de Tudhope, Koch, Heery (2006) e Shintaku, Gottschalg, Suaiden (2015) que enfatizam a ocorrência de problemas de recuperação ser causada pela falta de controle de vocabulário durante a indexação.

Tudhope, Koch, Heery (2006), realizaram pesquisa de revisão de literatura sobre serviços de terminologia e tecnologia para o Joint Information Systems Committee (JISC) do Reino Unido com o objetivo de orientar trabalhos futuros relacionados a Serviços de Terminologia e Tecnologia. No item de repositórios recomendaram estudos de avaliação de comportamento do usuário para atender requisitos de pesquisa e recuperação para que serviços de terminologia de assunto sejam agregados aos metadados. Consideram ainda que técnicas de indexação e classificação de assuntos podem sustentar vários serviços de repositórios tais como alerta personalizados ou coleta por conjunto de assuntos e recomendam diferentes abordagens para o acesso por assunto ao conteúdo do repositório mediante uso de diferentes tipos de vocabulário e serviços de terminologia conforme custo-benefício e níveis de agregação de conteúdo, tais como: classificação de assuntos, vocabulários controlados, KOS especializados, palavras-chave atribuídas pelo autor e indexação de texto completo.

Shintaku, Gottschalg, Suaiden (2015) apresentam considerações acerca de problemas relativos à indexação em repositórios a partir de investigação com técnicas de pesquisa documental em federações de repositórios, entre outros, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) ou a Networked Digital Library Theses and Dissertation (NDLTD. Verificaram três grupos de problemas da indexação que afetam a precisão e revocação na recuperação quanto à: qualidade dos dados, questões linguísticas e duplicidade de registros. No primeiro grupo, da qualidade, a falta de padronização de campos e normalização de conteúdo são fatores de problemas que impedem a plena indexação; no segundo grupo, de questões linguísticas, idiomas distintos e homonímias indicam que a indexação é apenas sintática sem controle de vocabulário para fazer a indexação semântica; e, no terceiro grupo, de duplicidade de registros que vem de fontes distintas e podem ser detectadas com ferramentas de indexação.

A análise dos textos demonstrou que investigadores ao longo dos anos, desde que surgiram as primeiras bases de dados, continuam realizando estudos exploratórios para observar comportamentos profissionais, de gestores e de usuários, ferramentas, processos, produtos e ambiente com diferentes contextos e tipologias textuais. Estudos de aplicação de novas metodologias, processos e produtos também são encontrados e demonstram novas possibilidades de soluções conciliadoras.

## 5 CONCLUSÕES

Ao fim desse ensaio sobre o debate entre linguagem natural e linguagem controlada, é possível deduzir sobre a existência de dois lados da moeda da representação que se completam mutuamente: o lado da representação na indexação, seja pelo autor ou pelo bibliotecário indexador, e o lado da representação na busca por assunto. Qualquer sistema de informação dotado, minimamente, de uma interface de busca, tem que pensar em agregar valor à busca por assunto, caso contrário, o sistema cai em desuso rapidamente. Nesse lado da moeda, está o usuário que, embora, seja desconhecido fisicamente, é conhecido virtualmente em muitos sistemas de busca a começar pelo Google cujas ferramentas baseadas em inteligência artificial muito bem desenvolvidas "aprendem" tudo sobre as preferências e não preferências de qualquer usuário a partir do comportamento de busca por assunto que, invariavelmente, será realizada com linguagem natural. Mas, sempre é possível realizar a "*tradução*" dessa linguagem natural e até corrigir eventuais erros de digitação agregando ferramentas de correção ortográfica (sintáticas), além de sugerir outros termos de busca semanticamente relacionados. Tudo isso faz parte de um processo de indexação que está "*por trás*" do sistema e é "*naturalmente*" incorporado sem que tenhamos que ser orientados para aprender a fazer buscas. Tudo é realizado intuitivamente aliado à nossa inteligência como se fosse a extensão de nossas mentes. Para isso, é necessário pensar na representação durante a indexação e incorporar ferramentas de controle da variação terminológica e vocabulários controlados para resolver problemas semânticos e ampliar as possibilidades de busca com a combinação de palavras-chave da linguagem natural e descritores da linguagem controlada.

A análise da revisão de literatura revela que os sistemas de informação devem aprender sobre indexação e a necessidade de tradução com uso de vocabulários controlados se quiserem ter sucesso em sistemas de buscas por assunto. A questão não é somente sobre encontrar alguma informação, mas encontrar a informação mais relevantemente relacionada à busca acompanhada de todas as outras informações armazenadas no sistema de informação e, para isso, é preciso empregar instrumentos, métodos e técnicas para o alcance da precisão terminológica.

Sistemas de busca de sistemas de informação precisam de desenvolvimento e padronização e na revisão de literatura é possível verificar quatro tipos de **soluções** interessantes e passíveis de serem testadas e avaliadas não só internamente, mas, principalmente, pelos usuários:

- Uma das principais soluções oferecidas referem-se ao uso combinado de vocabulário controlado com pesquisa de palavras-chave;

- O uso dos termos de busca do usuário para vocabulários controlados é, sem dúvida, uma solução propícia ao aprimoramento de vocabulários controlados de áreas especializadas;
- Uso de sistemas de marcação do usuário para indexação social é outra solução oferecida como forma de aprimorar a representação na indexação e recuperação;
- Uso combinado das linguagens natural e controlada como outra solução oferecida;

A reflexão principal aqui passa pela situação de que, atualmente, além de utilizar a linguagem natural na busca, o usuário poderá ser também o autor que faz a submissão de sua produção científica utilizando a linguagem natural para atribuir palavras-chave na indexação durante o preenchimento do metadado assunto. Esse autor/indexador/usuário não está acostumado ou orientado a usar um vocabulário controlado e a linguagem natural parece ser a alternativa mais adequada tendo em vista que ele é o especialista no assunto não fosse pela necessidade de as publicações científicas serem recuperadas e alcançarem o maior número de leitores para que sejam muito citadas. Outro ponto a ser considerado é que nem sempre existe vocabulário controlado indicado para uso alternativo à linguagem natural (GONÇALVES, 2008) ou mesmo a possibilidade de o sistema ter o controle de vocabulário a partir de mapeamento de vários vocabulários controlados.

Em estudo sobre análise das palavras-chave selecionadas pelos autores dos artigos em periódico especializado em Fotografia, Rodrigues, et al. (2017) percebem falta de reflexão por parte do autor em sua escolha de palavras-chave. Garcia, Gattaz e Gattaz (2019, p.6) alertam os autores quanto à escolha de palavras-chave na submissão de artigos e se referem a elas como "porta de acesso ao texto" e caso subestimem essa etapa, tal conduta poderá acarretar a "[...] sedimentação de seus textos na grande base submersa do iceberg das publicações". Esse alerta serve aos gestores de periódicos e repositórios que precisam pensar na visibilidade métrica das publicações.

Portanto, a conclusão do ensaio sobre o debate entre linguagem natural e linguagem controlada é pela coexistência para a representação na indexação e na recuperação com suporte de soluções que permitam a combinação de ambas dada a atualização contínua da linguagem natural e por ser a linguagem do usuário. Recomenda-se o desenvolvimento de pesquisas de aprimoramento de vocabulários controlados especializados com inclusão contínua de novos termos da linguagem natural, bem como do aprimoramento dos sistemas de busca com uso de técnicas intuitivas a partir de controle de vocabulário em nível sintático, semântico e pragmático.

## REFERÊNCIAS

AQUINO, I.S.; AQUINO, I.S. Análise sobre a forma da escrita de palavras-chave em artigos científicos na área de ciências agrárias publicados no período de 1999 a 2011. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, v. 18, n. 37, p. 227-238, mai./ago., 2013. DOI: 10.5007/1518-2924.2013v18n37p227

ARMANI B., et al. Repository management in an intelligent indexing approach for multimedia digital libraries. In: Raś Z.W., Ohsuga S. (eds) **Foundations of Intelligent Systems**. ISMIS 2000. Lecture Notes in Computer Science, v. 1932. Springer, Berlin, Heidelberg, 2000. https://doi.org/10.1007/3-540-39963-1_8

BACHA, M.N.; ALMEIDA, M.do S.G. de Vocabulário controlado e palavras-chave em repositórios digitais: relato de experiência do repositório institucional da FGV. XXV Congresso Brasileiro de Biblioteconomia, Documento e Ciência da Informação, Florianópolis, SC, Brasil, 07 a 10 de julho de 2013. **Anais...** Florianópolis: UFSC, 2013. p.1-8

BARITÉ, M. El control de vocabulario en la era digital: revisión conceptual. **Scire**, v,20, n.1, en.-jun. 2014.

BORST, T. Usage and impact of controlled vocabularies in a subject repository for indexing and retrieval. **Liber Quarterly**, v.21, n.3/4, p.445-53, 2012. DOI: http://doi.org/10.18352/lq.8035

CHU, H. **Information representation and retrieval in the digital age**. Medford, NJ: Information Today, 2003. 248p. (ASIST Monograph Series)

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

FREITAS, M. P. de **Autoarquivamento e representação de assunto**: estudo analítico de teses e dissertações do Repositório Institucional da UFSCar / Marina Penteado de Freitas. -- 2019. 89 f. : 30 cm. Dissertação (mestrado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos. https://repositorio.ufscar.br/handle/ufscar/11850

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*, v.23, n.3, maio/junho, 2019. Disponível em: www.scielo.br/pdf/rac/v23n3/1982-7849-rac-2019190178.pdf. Acessado 30 jun. 2020.

GOLUB, K. Potential and challenges of subject access in libraries today on the example of swedish libraries. **International Information & Library Review**, v.48, n.3, p.204-10, 2016. DOI: 10.1080/10572317.2016.1205406

GOLUB, K. Subject access in Swedish discovery services. **Knowledge Organization**, v.45, n.4, p.297-309, 2018. Disponível em: https://doi.org/10.5771/0943-7444-2018-4-297.

GOLUB, K., LYKKE, M., TUDHOPE, D. Enhancing social tagging with automated keywords from the Dewey Decimal Classification. **Journal of Documentation**, v.70, n.5, p.801-28, 2014. Disponível em: https://doi.org/10.1108/JD-05-2013-0056.

GOLUB, K., et al. Subject indexing in humanities: a comparison between a local university repository and an international bibliographic service. **Journal of Documentation**, v. ahead-of-print, n. ahead-of-print, 2020. https://doi.org/10.1108/JD-12-2019-0231

GOMES, H.E. **Classificação, tesauro e terminologia**: fundamentos comuns. Biblioteconomia, Informação & Tecnologia da Informação, S.d. Disponível em: http://www.conexaorio.com/biti/tertulia/tertulia.htm

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.

GROSS, T., TAYLOR, A.G What have we got to lose? The effect of controlled vocabulary on keyword searching results. **College & Research Libraries**, v. 66, n. 3, p. 212-230, may 2005. Disponível em: https://crl.acrl.org/index.php/crl/article/view/15726. Acesso em: 02 sep. 2020. doi:https://doi.org/10.5860/crl.66.3.212.

GROSS, T., TAYLOR, A.G., JOUDREY, D.N. Still a lot to lose: the role of controlled vocabulary in keyword searching. **Cataloging & Classification Quarterly**, v.53, n.1, p.1-39, 2015. DOI: 10.1080/01639374.2014.917447

HAN, M.-J. K., et al. (2016). Aligning author-supplied keywords for ETDS with domain-specific controlled vocabularies. In: CLASSIFICATION & INDEXING SATELLITE CONFERENCE, 2016 (pp. 1-10). Recuperado de http://hdl.handle.net/2142/97879

HANRATH, S.; RADIO, E. User search terms and controlled subject vocabularies in an institutional repository. **Library Hi Tech**, v.35, n.3, pp. 360-367, 2017.

HARTLEY, J., KOSTOFF, R.N. How useful are 'key words' in scientific journals? **Journal of Information Science**, v.29, n.5, p.433-438, 2003. Disponível em: journals-sagepub-com.ez78.periodicos.capes.gov.br/doi/pdf/10.1177/01655515030295008. Acessado em 17 jul. 2020.

HENZLER, R.G. Free or controlled vocabularies: some statistical user-oriented evaluations of biomedical information systems. **International Classification**, v.5, n.1, p.21-26, 1978. Acessível em: https://www.ergon-verlag.de/isko_ko/downloads/ic_5_1978_1_e.pdf

KIPP, M. E. I.. Complementary or discrete contexts in online indexing: a comparison of user, creator, and intermediary keywords. **Canadian Journal of Information and Library Science**, v.29, n.4, p.419-436, 2006.

KHOO, M., et al. Towards digital repository interoperability: the document indexing and semantic tagging interface for libraries (DISTIL). In: Zaphiris P., Buchanan G., Rasmussen E., Loizides F. (eds) **Theory and Practice of Digital Libraries**. TPDL 2012. Lecture Notes in Computer Science, v. 7489. Springer, Berlin, Heidelberg, 2012. https://doi.org/10.1007/978-3-642-33290-6_49.

LANCASTER, F.W. El control del vocabulario en la recuperación de información. 2.ed. Valencia: Universitá de Valencia, 2002. 286p.

LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ci. Inf.,** Brasília , v. 31, n. 1, p. 41-52, jan. 2002. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100005&lng=pt&nrm=iso. acessos em 18 ago. 2020.

LU, W. et al. How do author-selected keywords function semantically in scientific manuscripts? **Knowledge Organization,** v.46, n.6, p.403-18, 2019. https://doi.org/10.5771/0943-7444-2019-6-402

MATERSKA, K. Knowledge organization in university repositories in Poland. In: ISKO International, 2016. **Advances in Knowledge Organization**, v.15, p.69, 2016

MAURER, M.B., SHAKERI, S. Disciplinary differences: LCSH and keyword assignment for ETDs from different disciplines. **Cataloging & Classification Quarterly**, v.54, n.4, p.213-243, 2016. DOI: 10.1080/01639374.2016.1141133

MEDEIROS, G.M. de **Organização da informação em repositórios digitais:** implicações do auto-arquivamento na representação da informação**.** 2010**.** Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-graduação em Ciência da Informação, Florianópolis, 2010

MIGUÉIS, A. et al. A importância das palavras-chave dos artigos científicos da área das Ciências Farmacêuticas, depositados no Estudo Geral: estudo comparativo com os termos atribuídos na MEDLINE. **InCID**: Revista de Ciência da Informação e Documentação, vol. 4, no. 2, dez. 2013, p. 112-125, Disponível em: www.revistas.usp.br/incid/article/view/69284. Acessado 16 jun. 2020.

MIGUÉIS, A.; NEVES, B. Uma abordagem à linguagem de indexação dos artigos científicos depositados no repositório científico da Universidade de Coimbra. **Ponto de Acesso**, v.7, n.1, p.116-31, 2013. http://hdl.handle.net/10316/23450

PAES, M.L. **Arquivo**: teoria e prática. 3.ed.rev.ampl. Rio de Janeiro: Fundação Getúlio Vargas, 2004. 228p.

RODRIGUES, M.R. et al. Tratamento temático da informação na revista discursos fotográficos: palavras-chave ou descritores? In: SEMINÁRIO EM CIÊNCIA DA INFORMAÇÃO - SECIN, 7., 2017. **Anais eletrônicos** […], UEL, 2017. p. 1063-1076, www.uel.br/eventos/cinf/index.php/secin2017/secin2107/schedConf/presentations. Acessado 28 out. 2018.

ROWELL, C.J. **Controlled Vocabulary Use by Data Repositories**: Determining Status and Potential for Promoting Interoperability. A Master's paper for the M.S. in Information Science degree. July, 2013. 65 p. Advisor: Jane Greenberg

SANTARÉM SEGUNDO, J. E. **Representação iterativa**: um modelo para repositórios digitais. Marília, 2010. Tese (Doutorado em Ciência da Informação) –Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, 2010.

SANTOS, Fatima Cristina Lopes dos. **Coerência na representação temática de artigos científicos indexados no repositório de saúde pública da Fundação Oswaldo Cruz**. 2017. 258 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal do Rio de Janeiro/Escola de

Comunicação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Programa de Pós-Graduação em Ciência da Informação, Rio de Janeiro, 2017.

SANTOS, R. F. DOS. Indexação em repositórios digitais: uma abordagem sobre o metadado assunto da Biblioteca Digital de Monografias da UFRN. **Revista Informação na Sociedade Contemporânea**, v. 1, p. 1-22, 11 jun. 2017. Disponível em:https://periodicos.ufrn.br/informacao/article/view/12279

SANTOS, R. F. **Modelos colaborativos de indexação social e a sua aplicabilidade na base de dados referencial de artigos de periódicos em ciência da informação (BRAPCI).** 184f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, Recife, 2016.

SANTOS, R. F. dos; CORRÊA, R. Modelos colaborativos de indexação social e sua aplicabilidade em bibliotecas digitais. **Liinc em Revista**, v.11, n.1, p.273-86, 2015. Disponível em: doi: http://dx.doi.org/10.18225/liinc.v11i1.768

SANTOS, R.F. dos; NEVES, D.A.de B. Práticas de indexação em repositórios digitais de acesso aberto: análise do metadado do assunto do Repositório Instittucional da UFRN. In: NEVES, D.A. de B.; SANTOS, R.F. dos S.; GUIMARÃES, I.J.B. **Práticas e reflexões sobre a representação da informação em cenários informacionais.** São Leopoldo: Karywa, 2019. p.49-64.

SASSEN, C. Enhancing bibliographic access to dissertations. **Technical Services Quarterly**, v.34, n.1, p.40-53, 2017. DOI: 10.1080/07317131.2017.1238202

SHINTAKU, M; GOTTSCHALG, C.D.; SUAIDEN, E.J. Federações de repositórios: conceitos, políticas, características e tendências. **Perspectivas em Ciência da Informação**, v.20, n.3, p.51-66, jul. /set. 2015. Acessível em: http://dx.doi.org/10.1590/1981-5344/2358

SILVA, M.F. **Proposta de modelo de colaboração para catálogo web facetado**. Belo Horizonte, 2013. 269f. Tese (Dou. em C. da Inf.). Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

SILVA, M.F.; LIMA, G.Â.B. de O. Avaliação de usabilidade em interface de busca com navegação facetada e busca por palavra-chave. **Tendências da Pesquisa Brasileira em Ciência da Informação**, n. 1, v. 8, 2015. Disponível em: https://brapci.inf.br/index.php/res/v/119572. Acesso em: 06-ago.-2020.

SMIRAGLIA, R.P. Keywords Redux – an editorial. **Knowoledge Organization***,* v.42, n.1, p.4-8, 2015.

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

SOUZA, M.I.F. et al. Representação descritiva e temática no Sistema Agência de Informação Embrapa: controle de vocabulário. **Transinformação**, v.22, n.1, p.61-75, 2010.

TARTAROTTI, R.C.D. **Avaliação do processo de indexação de assuntos em repositórios institucionais pela abordagem da recuperação da informação**. 370p. Tese (Doutorado) - Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp). Faculdade de Filosofia e Ciências. Programa de Pós-graduação em Ciência da Informação. Marília, 2019.

TUDHOPE, D.; KOCH, T.; HEERY, R. **JISC state of the art review**. Bath, Somerset, Reino Unido: UKOLN, 2006.

VIANA, J.M.dos A. **Representação colaborativa de dados científicos**: estudo na rede de repositórios de dados científicos do Estado de São Paulo / Joyce Mirella dos Anjos Viana. -- 2020. 127 f. : 30 cm. Dissertação (mestrado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos. https://repositorio.ufscar.br/handle/ufscar/13027

WHITE, H. Examining scientific vocabulary: mapping controlled vocabularies with free text keywords. **Cataloging & Classification Quarterly**, v.51, n.6, p.655-674, 2013. DOI: 10.1080/01639374.2013.777004

ZAVALINA, O. L. Complementarity in subject metadata in large-scale digital libraries: a comparative analysis, **Cataloging & Classification Quarterly**, v.52, n.1, p.77-89, 2014. DOI: 10.1080/01639374.2013.848316

ZHANG, Y, et al. Controlled vocabularies for scientific data: users and desired functionalities. **ASIST 2015***,* nov. p. 6-10, 2015, St. Louis, MO, USA.