UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
FACULDADE DE FILOSOFIA E CIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JOSÉ CARLOS FRANCISCO DOS SANTOS

INTEROPERABILIDADE DE VOCABULÁRIOS CONTROLADOS EM PERIÓDICOS CIENTÍFICOS ELETRÔNICOS: um estudo de caso de compatibilização sistemática por meio dos padrões de Hearst

MARÍLIA, SP
2020

JOSÉ CARLOS FRANCISCO DOS SANTOS

INTEROPERABILIDADE DE VOCABULÁRIOS CONTROLADOS EM PERIÓDICOS CIENTÍFICOS ELETRÔNICOS: um estudo de caso de compatibilização sistemática por meio dos padrões de Hearst

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências da Universidade Estadual Paulista - UNESP, como requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

Orientador: Prof. Dr. Walter Moreira

Linha de Pesquisa: Produção e Organização da Informação

MARÍLIA, SP
2020

S237i

Santos, José Carlos Francisco dos

Interoperabilidade de vocabulários controlados em periódicos científicos eletrônicos : um estudo de caso de compatibilização sistemática por meio dos padrões de Hearst / José Carlos Francisco dos Santos. -- Marília, 2020

163 p. : il.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília

Orientador: Walter Moreira

1. Sistema de Organização do Conhecimento. 2. Vocabulário Controlado. 3. Mapeamento. 4. Compatibilidade. 5. Modelo de Interoperabilidade. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

JOSÉ CARLOS FRANCISCO DOS SANTOS

INTEROPERABILIDADE DE VOCABULÁRIOS CONTROLADOS EM PERIÓDICOS CIENTÍFICOS ELETRÔNICOS: um estudo de caso de compatibilização sistemática por meio dos padrões de Hearst.

BANCA EXAMINADORA

Prof. Dr. Walter Moreira (Orientador)
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Profª. Dra. Mariângela Spotti Lopes Fujita
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Profª. Dra. Deise Maria Antonio Sabbag
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Profª. Dra. Brígida Maria Nogueira Cervantes
Departamento de Ciência da Informação
Universidade Estadual de Londrina (UEL)

Profª. Dra. Cibele Araújo Camargo Marques dos Santos
Escola de Comunicações e Artes (ECA)
Universidade de São Paulo (USP)

Marília – SP, 17 de março de 2020.

AGRADECIMENTOS

A Deus por toda inspiração divina que contribuiu de forma significativa durante a realização do Programa de Doutorado, desenvolvimento e construção da tese, que enfim resulta em mais uma meta alcançada.

Ao professor Dr. Walter Moreira pela atenção, atitude de parceria, apoio constante, pelos conhecimentos compartilhados com toda a dedicação, pelos apontamentos precisos em cada detalhe que contribuísse para o aperfeiçoamento da tese. Foi um privilégio estar sob sua supervisão.

À professora Brígida Maria Nogueira Cervantes, a quem dedico um afeto semelhante ao filial, minha gratidão pelas contribuições e aconselhamento sobre caminhos a percorrer, por me indicar e me incentivar nos passos necessários à consolidação do meu percurso de pesquisador, e também por me fazer descobrir o fascínio de trabalhar com o vocabulário controlado.

À professora Mariângela Spotti Lopes Fujita pelo acolhimento, pela maneira gentil de analisar a minha produção, pelas recomendações tanto técnicas quanto metodológicas, sem dúvida uma contribuição essencial na temática de vocabulário controlado, sob os auspícios de seu grupo de trabalho.

Ao professor Miguel Contani, pelas contribuições no desenvolvimento, organização e revisão textual e, em especial, por sua dedicação e orientação durante toda minha vida acadêmica.

A todo o corpo docente do Programa de Pós-Graduação em Ciência da Informação – PPGCI pelo conhecimento novo que me permitiram adquirir e pelo crescimento que me proporcionaram em todas as oportunidades de contato direto, em especial ao Prof. José Augusto Guimarães e Prof. Edberto Ferneda.

À Profª Marta Ligia Pomim Valentim coordenadora do PPGCI Unesp, que sempre trouxe orientações quanto ao direcionamento da pesquisa, publicações, convivência com os demais colegas do PPGCI ao longo dos eventos.

Aos membros das bancas de qualificação, pelas indicações que serviram de alicerce às investigações realizadas. Também pelas contribuições recebidas em outras oportunidades de reunião, grupos de pesquisa e grupos de trabalho.

Aos amigos da turma PPGCI, pelo inesquecível convívio e pelo que compartilhamos de muitos momentos importantes de socialização do conhecimento, em especial àqueles com a mesma filiação acadêmica, nos grupos de pesquisa, disciplinas, publicações.

Enfim, são tantas as pessoas que colaboraram no desenvolvimento desta pesquisa, de forma direta e indireta, que se torna impossível nominá-las, sem correr o risco de omitir alguém, de modo que desejo que se sintam todas representadas neste meu agradecimento especial.

Dedico este trabalho

À minha família fonte de inspiração, pelo apoio, paciência nos momentos de difíceis e amor, durante todo esse período de estudos e ausências, minha esposa Fabiana, meu filho Antonio Carlos e meu filho José Eduardo.

Aos meus pais Sebastião e Terezinha, pelo apoio de sempre quando diz respeito à ampliação dos conhecimentos.

Aos meus irmãos Bruna, Paulo, Altair e Luzia, cunhados, cunhadas, sobrinhos, sogro, sogra, e demais familiares que me apoiaram durante esta etapa.

Aos colegas do ESAP pelo apoio e compreensão nas minhas ausências.

Aos colegas de trabalho e docentes das Faculdades Integradas do Vale do Ivaí – Univale pelo apoio nesta e nas demais etapas acadêmicas.

À professora Neila Estigarribia e Sr. Hypólito Haluch (in memorian), diretores do ESAP e Faculdades Integradas do Vale do Ivaí, por todos os ensinamentos profissionais e técnicos e a flexibilização das minhas atividades laborais para o desenvolvimento da pesquisa.

A sabedoria, porém, que vem de cima, é primeiramente pura, depois pacífica, condescendente, conciliadora, cheia de misericórdia e de bons frutos, sem parcialidade, nem fingimento.
São Tiago 3,17

SANTOS, José Carlos Francisco dos. **Interoperabilidade de vocabulários controlados em periódicos científicos eletrônicos:** um estudo de caso de compatibilização sistemática por meio dos padrões de Hearst. 2020. 163 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências da Universidade Estadual Paulista – UNESP, Marília, 2020.

RESUMO

O controle de vocabulário em periódicos científicos eletrônicos tem como finalidade minimizar ou extinguir a ambiguidade da linguagem natural para proporcionar uma linguagem artificial com vista a organização, representação e recuperação da informação. Os vocabulários controlados originados nos periódicos científicos eletrônicos deste estudo de caso são disponibilizados por meio da ferramenta VCPC *Tools*, e constituídos independentes, o que instiga a necessidade de mapear e torná-los interoperáveis. A contribuição pretendida com a realização deste estudo é a ampliação das condições de realizar inferências sobre a interoperabilidade entre Sistemas de Organização do Conhecimento e aplicá-las em periódicos científicos eletrônicos gerenciados pelo OJS que já fazem uso de vocabulário controlado. Busca-se responder a questão: como recuperar a informação em periódicos científicos eletrônicos por meio de vocabulários controlados mapeados a partir das palavras-chave em um modelo de interoperabilidade? Partiu-se da hipótese inicial de que os vocabulários controlados aplicados a periódicos científicos eletrônicos, gerenciados pelo OJS, podem ser considerados um instrumento interoperável para melhorar os processos de busca e recuperação dos artigos, bem como subsidiar os autores (no momento de atribuir) e os usuários na operação de encontrar, na consulta por meio de termos que representem, de modo mais eficaz, a sua pesquisa. Defende-se a tese de que o controle de vocabulário interoperável, num contexto de grande volume de produções científicas, é de extrema importância para o acompanhamento desse crescimento científico, e não prescinde da produção de instrumentos para tratamento e aperfeiçoamento das formas de recuperação da informação. O objetivo geral é apresentar uma proposta teórico-metodológica (um modelo) de interoperabilidade entre vocabulários controlados de periódicos científicos eletrônicos. A metodologia é integrada em duas etapas: fundamentos teóricos (descritiva) e estudo de caso para o projeto experimental. O vocabulário controlado interoperável denominado VCPC-CI é o instrumento (produto) resultante desta tese, concebido a partir da operação de compatibilização sistemática das palavras-chave por meio dos padrões de Hearst e do mapeamento reverso das palavras-chave compatibilizadas por Santos (2015), totalizando 926 termos. O protótipo da interface de busca faz a integração entre o VCPC-CI gerenciado pelo TemaTres e o OJS, utilizando como suporte o VCPC *Tools*. Conclui-se que os periódicos científicos eletrônicos ainda carecem de instrumentos de controle de vocabulário para organizar e representar o conteúdo temático dos artigos e, consequentemente, recuperá-los de forma significativa. A partir da proposta teórico-metodológica desenvolvida e aplicada, é possível sustentar a viabilidade de construção de vocabulários controlados interoperáveis como recurso eficaz para uso como parte integrante de periódicos científicos eletrônicos.

Palavras-chave: Sistema de Organização do Conhecimento. Vocabulário Controlado. Mapeamento. Compatibilidade. Modelo de Interoperabilidade. Periódico Científico Eletrônico. VCPC-CI.

SANTOS, José Carlos Francisco dos. **Interoperability of controlled vocabularies in electronic scientific journals:** a case study of systematic compatibility through Hearst standards. 2020. 163 f. Thesis (PhD Degree in Information Science) - Faculdade de Filosofia e Ciências da Universidade Estadual Paulista – UNESP, Marília, 2020.

ABSTRACT

Vocabulary control in electronic scientific journals aims to minimize or extinguish the ambiguity of natural language, in order to provide an artificial language heading towards organizing, representing and retrieving information. The controlled vocabularies originated in the electronic scientific journals, in this case study, are made available through VCPC *Tools*, as independently set device, which instigates the need to map and make them interoperable. The intended contribution, with the accomplishment of this study, is to widely enable users to make inferences about the interoperability between SOCs – Knowledge Organization Systems, and to apply them in electronic scientific journals managed by the OJS – Open Journal Systems, that already utilize controlled vocabulary. The question to be answered is: how to retrieve information in electronic scientific journals through controlled vocabularies mapped from keywords in an interoperability model? The initial hypothesis is that the controlled vocabularies when applied to electronic scientific journals, managed by the OJS, can be considered as an interoperable instrument to improve the processes of searching and retrieving articles, and constitute the basis for the authors (when assigning keywords) and for users (in the operation of finding) to properly search, through more effective terms and refined access. The thesis being held is that interoperable controlled vocabulary, in a context of a huge amount of scientific productions, is extremely important keep pace with scientific growth, and cannot give up the production of instruments to handle and improve the ways to recover information. The general objective is to present a theoretical-methodological proposal (a model) of interoperability between controlled vocabularies of electronic scientific journals. The methodology is integrated into two stages: theoretical foundations (exploratory and descriptive) and case study for the experimental project. The interoperable controlled vocabulary called VCPC-CI is the instrument (product) resulting of this thesis, and is part of the systematic compatibility of keywords through Hearst standards, and of the reverse mapping of keywords by Santos (2015) including 926 terms. The prototype of the search interface integrates between VCPC-CI managed by TemaTres and OJS, using VCPC *Tools* as support. Major findings point out that electronic scientific journals still lack vocabulary control instruments to organize and represent the thematic content of the articles and, consequently, recall them properly. As a result, based on the theoretical-methodological proposal developed and applied, it is possible to sustain the feasibility of building interoperable controlled vocabularies as an effective resource to use as an integral part of electronic scientific journals.

Keywords: Knowledge Organization System. Controlled Vocabulary. Mapping. Compatibility. Interoperability Model. Electronic Scientific Journal. VCPC-CI.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
BARTOC	*Basel Registerof Thesauri, Ontologies & Classifications*
BDTD	Biblioteca Digital Brasileira de Teses e Dissertações
CECA	Centro de Educação, Comunicação e Artes
CI	Ciência da Informação
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DDC	*Dewey Decimal Classification*
DeCS	Descritores em Ciência da Saúde
EIES	*Eletronic Information Echange System*
FEBAB	Federação Brasileira de Associações de Bibliotecários e Instituições
HTML	*HyperText Markup Language*
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
ICD	*International Classification of Diseases*
IES	Instituição de Ensino Superior
IFLA	*International Federation of Library Associations and Institutions*
INPI	Instituto Nacional da Propriedade Industrial
ISTA	*Information Science and Technology Abstracts*
KDS	*Clinical Key of Diagnosis*
LA	Linguística Aplicada
LC	Linguística Computacional
LCC	*Library of Congress Classification*
LISA	*Library Information Science Abstratcs*
LISTA	*Library and Informations Science Abstracts*
LISTA	*Library, Information Science & Technology Abstracts*
LOB	*Lancaster-Oslo/Bergen Corpus*
LOD	*Linked Open Data*
LSF	Linguística Sistêmico-Funcional
LSPE	*Lexico-Syntactic Pattern Extraction*
MeSH	*Medical Subject Headings*
OC	Organização do Conhecimento
ODS	Objetivos de Desenvolvimento Sustentável
OJS	*Open Journal Systems*
ORC	Organização e Representação do Conhecimento
PDF	*Portable Document Format*
PKP	*Public Knowledge Project*

RDF	*Resource Description Framework*
SEER	Sistema de Editoração Eletrônica de Revistas
SKOS	*Simple Knowledge Organization System*
SNOP	*Systematized Nomenclature of Pathology*
SOC	Sistemas de Organização do Conhecimento
TBCI	Tesauro Brasileiro de Ciência da Informação
TCI	Tesauro em Ciência da Informação
TI	Tecnologia da Informação
UEL	Universidade Estadual de Londrina
UML	*Unified Modeling Language*
VCPC-CI	Vocabulário Controlado dos Periódicos Científicos de Comunicação e Informação
W3C	*World Wide Web Consortium*

SUMÁRIO

1 INTRODUÇÃO

A organização e representação do conhecimento para a recuperação da informação sempre foi e será um desafio muito abrangente e complexo, tanto no campo da Ciência da Informação (CI) quanto na Ciência da Computação. Envolve processos de representação da informação, que vão ao encontro da necessidade de informação do usuário no momento da busca. Por um lado, a Ciência da Computação busca aperfeiçoar modelos e algoritmos envolvidos no processo de recuperação de informação, ou seja, cuida da efetividade destes sistemas no que diz respeito à procura do resultado mais relevante. Por outro lado, a CI tem suas atividades centradas na organização e representação da informação, com vistas à recuperação de informação e, neste caso, a avaliação do usuário sobre a qualidade e relevância do que é recuperado ganha destaque.

Com o surgimento dos periódicos científicos eletrônicos e da *Web*, o desafio de organizar e representar o conhecimento visando à recuperação da informação torna-se ainda maior diante da quantidade de informação e da complexidade envolvida. Os primeiros periódicos surgiram no século XVII, entre os anos de 1650 e 1700, de maneira impressa; os periódicos eletrônicos tiveram os primeiros experimentos de 1978 a 1980. No intervalo de tempo entre as aparições eletrônicas e o atual momento, as publicações científicas cresceram exponencialmente, fenômeno aliado ao contexto tecnológico em contínuo avanço, que contribui para a veloz disseminação de informações científicas.

Dentre as tecnologias, as redes de computadores, a *Web*, os sistemas de gerenciamento eletrônico de periódicos e demais ferramentas tecnológicas colaboram para acelerar este crescimento. O *Open Journal Systems* (OJS) é um sistema de editoração eletrônica de periódicos que foi desenvolvido pela *Public Knowledge Project* (PKP) e foi customizado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), em 2003, quando recebeu a denominação de Sistema de Editoração Eletrônica de Revistas (SEER). O início da distribuição foi em 2004 e, com base no histórico do IBICT até 2009, cinco anos de existência do SEER, 800 periódicos brasileiros já o utilizavam. Contribuiu, portanto, significativamente para o aumento das publicações científicas, especialmente os periódicos científicos eletrônicos.

Neste contexto, de uma expressiva quantidade de artigos científicos eletrônicos, amplia-se a complexidade de organização e representação do conhecimento, consequentemente dos processos de busca e recuperação da informação, motivo este que advém da implementação de instrumento para o controle de vocabulário como mecanismo de auxílio para representar os artigos científicos. A demanda por pesquisas relacionadas aos processos de representar

tematicamente os artigos científicos dos periódicos científicos eletrônicos gerenciados pelo OJS é atual e necessita de estudos com o foco nesta representação, tendo em vista a grande quantidade de publicações científicas e o crescimento contínuo deste número.

Desde a infância, o humano apresenta um comportamento classificatório nas atividades lúdicas, escolares, interação familiar e social, ao fazer classificação de objetos por meio de padrões de similaridades e distanciamentos. Esses aspectos remetem à organização do conhecimento a partir da iniciação interacional do humano em seus primeiros anos de vida.

As atividades de classificação são base das estruturas complexas de Sistemas de Organização do Conhecimento (SOC). De maneira geral, é muito abrangente o termo SOC, portanto é possível determinar que as linguagens criadas para indexar e classificar são instrumentos para realizar o tratamento de textos (BARITÉ, 2011). Estes instrumentos construídos para a organização do conhecimento objetivam a representação do conhecimento por meio dos SOC. Dois processos são fundamentais para a análise de assunto do documento e atribuição de conceitos. O primeiro resulta numa expressão linguística, já no segundo aplica-se uma linguagem documentária, a fim de padronizar o processo de indexação (NOVELLINO, 1996).

O desenvolvimento dos SOC, de modo geral, proporcionou as evoluções das formas de representação da informação e do conhecimento, auxiliando assim o trabalho árduo do profissional da informação na indexação de materiais para proporcionar uma busca e recuperação de resultados mais significativos. Nessa perspectiva, os SOC representam uma denominação geral para instrumentos como taxonomias, esquemas de classificação, tesauros, listas de cabeçalhos, listas de termos autorizados, entre outros instrumentos que desempenham, dentre outras, a função de vocabulário controlado.

Os vocabulários controlados tendem a aumentar a consistência interna dos sistemas de informação, porém esbarram em questões técnicas e de padronizações, que reduzem as possibilidades de compatibilização entre diferentes sistemas. Os sistemas baseados em linguagem natural podem ser mais compatíveis que os sistemas com vocabulário controlado, considerando as questões de idioma, aspectos conceituais e normalização desses sistemas. Por exemplo, duas bases de dados de resumos científicos e citações bibliográficas poderiam ser unidas facilmente se existisse a compatibilidade técnica entre elas. Isso é decorrente da proximidade da linguagem científica e ainda a busca poderia ser de forma única. Nesta situação, supõe-se que cada base de dados utiliza um tesauro diferente e existe compatibilidade entre eles, sendo possível juntar os registros, porém não será possível a busca com uma estratégia

comum, pois os conceitos podem estar representados de formas diferentes (LANCASTER, 1995; LANCASTER; SMITH, 1983).

O vocabulário controlado, para Cunha e Cavalcanti (2008, p. 378), tem como finalidade uniformizar a forma de armazenagem para facilitar a recuperação. Isso é realizado por meio de termos atribuídos aos materiais junto aos sistemas de informação no momento da indexação, no momento da recuperação estes mesmos termos devem ser utilizados.

Desse modo, os vocabulários controlados têm suas vantagens e apresentam três funcionalidades fundamentais: reduzir a ambiguidade das semânticas, melhorar a consistência da representação do conteúdo e facilitar a realização de busca. Os sistemas com linguagem natural oferecem vantagens sobre os sistemas que utilizam linguagem controlada. O vocabulário controlado permite uma grande especificidade de recuperação, porém quanto mais específica for a busca, maiores serão os resultados nos sistemas com linguagem natural. O vocabulário controlado é mais prático, oferece ao usuário um ponto de busca e reduz a possibilidade da busca ser incompleta (LANCASTER, 1995; LANCASTER; SMITH, 1983).

O vocabulário controlado aliado à tecnologia soma-se como uma contribuição para o aperfeiçoamento da organização e representação do conhecimento, busca e recuperação da informação. Auxilia os profissionais da informação para a representação do contexto dos documentos, de maneira que seja a mais fidedigna do conteúdo. Trata-se de tarefa árdua e complexa, um grande desafio para os profissionais da área de Informação. Por outro lado, o usuário-pesquisador é beneficiado no momento de representar a sua busca por meio de um vocabulário controlado, para a possível recuperação de conteúdos informacionais temáticos que condizem com sua necessidade de busca.

Para efeito deste estudo foi adotado o termo vocabulário controlado para denominar o protótipo de instrumento interoperável que, futuramente, as organizações mantenedoras dos periódicos envolvidos no estudo de caso poderão nomear conforme a evolução da complexidade, dos relacionamentos dos termos e a consolidação da estrutura conceitual do vocabulário. Considera-se que o vocabulário controlado, conforme a definição da norma ISO 25964:1 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011), é uma lista prescritiva de termos, em que cada um deles representa um conceito. De modo sumário, os vocabulários controlados dos periódicos (Informação & Informação, Informação@Profissões e Discursos Fotográficos) estão disponibilizados como uma lista de termos controlados compatibilizados com o Tesauro Brasileiro de Ciência da Informação (TBCI) de Pinheiro e Ferrez (2014), de forma não ligada, observa-se que o TBCI era disponibilizado em PDF e em 2018 Santos, Cervantes e Fujita (2018) realizaram a importação

com o recurso do sistema TemaTres. Desta forma justifica-se a adoção destes termos para esta pesquisa.

Neste contexto, vale abordar a VCPC *Tools*[1] desenvolvida e aplicada por Santos (2015), apresentada na Figura 1. Na pesquisa (SANTOS, 2015), aborda-se o controle de vocabulário em periódicos científicos eletrônicos a partir das palavras-chave atribuídas aos artigos científicos do periódico científico eletrônico. A ferramenta VCPC *Tools* tem por finalidade a coleta de palavras-chave e a compatibilização dessas palavras-chave com um vocabulário controlado da área para almejar o controle de vocabulário e produzir um vocabulário controlado da própria revista relativo aos conteúdos publicados. O processo de desenvolvimento da ferramenta foi por meio da metodologia da engenharia de *software*, com destaque para o levantamento e modelagem de requisitos. Os requisitos da ferramenta são: RF01 Coleta de metadados; RF02 Análise de Metadados e Vocabulário controlado; RF03 Tratamento Palavras-chave; RF04 Resultados.

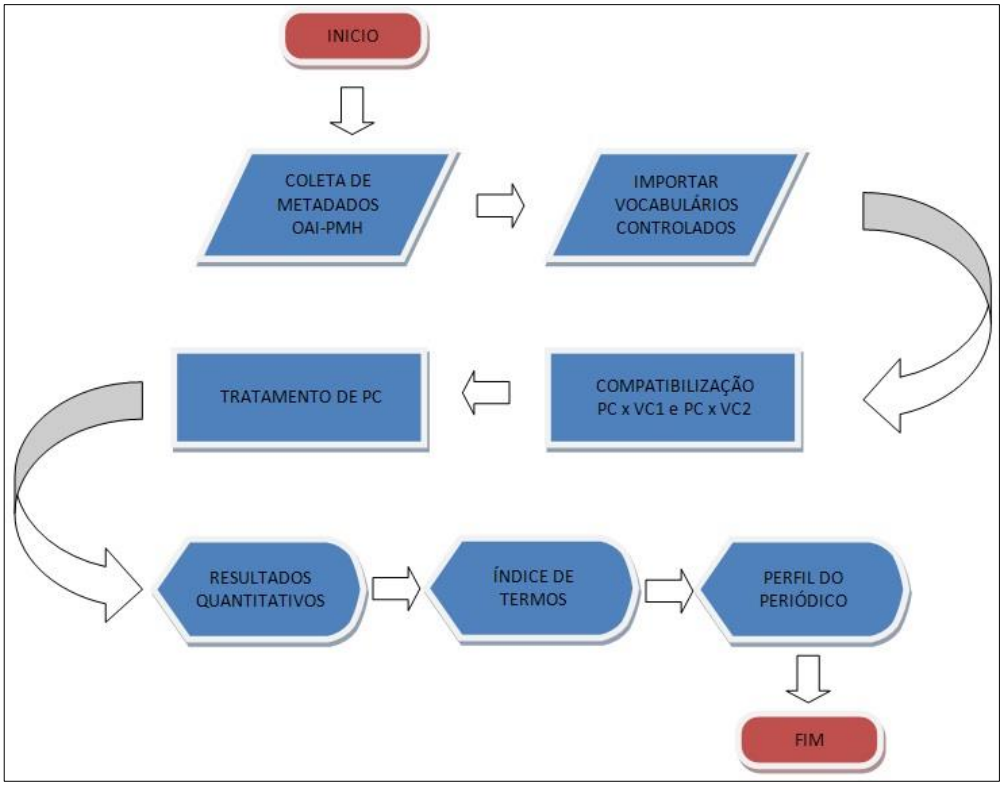
No RF01, realizou-se a coleta de metadados por meio do protocolo *OAI-PMH* do repositório digital do periódico, esses metadados foram tratados e armazenados na base de dados da VCPC *Tools*. A análise de metadados foi efetuada por meio dos mecanismos de comparação, denominada compatibilização das palavras-chave como o vocabulário controlado. São cinco processos de compatibilização, no primeiro comparou-se uma expressão da palavra-chave com cada termo do vocabulário controlado, denominada idêntica. No segundo retirou da palavra-chave as palavras vazias, fundamentada em parte dos processos do primeiro módulo da proposta de automatização da indexação de Gil Leiva (1999), a mesma regra foi aplicada para o termo do vocabulário controlado, a fim de obter a compatibilização. No terceiro aplicou-se a redução de plurais por meio do conjunto de regras e exceções de Orengo, Buriol e Coelho (2007), também aplicado no termo do vocabulário controlado. No quarto reduziu-se o sufixo por meio das regras e exceções de Orengo, Buriol e Coelho (2007). Por fim, no quinto calculou-se o índice contido entre cada palavra-chave e cada termo do vocabulário controlado, fundamentado no propósito de Gil Leiva (1999) de medir o índice de consistência.

Portanto o objetivo geral de Santos (2015) foi implantar o controle de vocabulário, tendo como base as palavras-chave atribuídas no periódico científico eletrônico Informação & Informação, para aprimorar a recuperação dos conteúdos temáticos. E os específicos foram:

[1] Nome do Software originado dos termos Vocabulário Controlado em Periódicos Científicos, desenvolvido como resultado da dissertação de Santos (2015), licença-patente emitida pelo Instituto Nacional da Propriedade Industrial (INPI), com as funcionalidades de coleta, compatibilização e disponibilização de palavras-chave de periódicos científicos eletrônicos gerenciados pelo OJS.

desenvolver a ferramenta VCPC *Tools*; coletar e armazenar os vocabulários controlados; compatibilizar as palavras-chave com os termos dos vocabulários controlados; tratar, intelectualmente, as palavras-chave não compatibilizadas automaticamente. Na pesquisa constatou-se a importância das palavras-chave como um elemento representativo do conteúdo do artigo científico eletrônico, que por sua vez é atribuído pelo autor, porém não tem sido dada a devida importância que é requerida.

Figura 1 - Sistematização VCPC *Tools*



Fonte: Santos (2015, p.21)

A VCPC *Tools* foi desenvolvida e aplicada para extrair as palavras-chave por meio da coleta de metadados dos artigos científicos eletrônicos, de forma a requisitar ao protocolo de interoperabilidade *OAI-PMH* e extrair do retorno desta requisição os metadados descritivos e sua inserção na base de dados. Na atividade de <importar vocabulários controlados> (Figura 1) é que se executam todos os procedimentos instanciados para coleta, tratamento, conversão e importação na base de dados da VCPC *Tools*. Todos esses procedimentos são necessários para trabalhar-se com vocabulários controlados que estão em formato não interoperáveis, ou seja, impressos, eletrônico estático (pdf), eletrônico na Web, entre outros suportes que não permitem a troca de informações.

O processo de compatibilização foi desenvolvido para executar as atividades de localização das palavras-chave dos artigos científicos eletrônicos, buscando a igualdade e a

similaridade com um termo do vocabulário controlado selecionado, neste caso, o TBCI. Na busca pela igualdade leva-se em consideração, literalmente, que o conjunto de caracteres da palavra-chave seja exatamente igual ao conjunto de caracteres do termo do vocabulário controlado. Na similaridade executam-se os processos de retirada de plurais, palavras vazias, sufixo e cálculo do índice de contingência do termo do vocabulário controlado na palavra-chave (SANTOS, 2015).

Na atribuição de palavras-chave em um artigo, o vocabulário controlado é de fundamental importância, significando um instrumento de auxílio para a representação do conteúdo informacional do artigo. A atribuição de palavras-chave aos artigos é, geralmente, realizada de forma livre e arbitrária pelo autor do trabalho, isso contribui para a ocorrência de imprecisões na recuperação da informação nos sistemas de informação documentária. Nesse contexto, insere-se o vocabulário controlado para eximir-se dessas inconsistências nos resultados de busca. Disso decorre a necessidade de aplicação de ferramentas tecnológicas - *software* - para a gerência de termos de instrumentos de controle de vocabulário integrado com interfaces de buscas de periódicos científicos eletrônicos.

A VCPC *Tools* vem sendo estudada para o desenvolvimento de soluções de integração com *software* gerenciador de vocabulários controlados, neste caso, o *TemaTres* (SANTOS; CERVANTES; LONDERO, 2018; SANTOS; CERVANTES; LONDERO; GONÇALEZ, 2016). Os vocabulários controlados disponibilizados pelos periódicos científicos eletrônicos e gerenciados pelo OJS necessitam de soluções de interoperabilidade. Nesse caso, refere-se à interoperabilidade dos vocabulários controlados de referência e os derivados, denominados vocabulário controlado do periódico científico, é resultante da compatibilização de palavras-chave atribuídas aos artigos científicos pelos autores com um vocabulário controlado de referência, tornando-as em um descritor, um termo ou um ponto de acesso em caso de compatibilização.

Insere-se, por esse motivo, nos processos de compatibilização, a compatibilização sistemática, assim denominada neste estudo, a qual será realizada por meio da identificação dos padrões de Hearst (1992, 1998), que são padrões léxicos-sintáticos com a finalidade de extração de relações lexicais hiponímicas (MOREIRA; SANTOS; VITORINI, 2017), e consistirá como suporte para a compatibilização de palavras-chave a partir de termos gerais e/ou específicos com a estrutura sistemática do vocabulário controlado. Vale registrar que o estudo de Moreira, Santos e Vitorini (2017), desenvolvido no Programa de Pós-graduação da Universidade Estadual Paulista (UNESP), é pioneiro na aplicação dos padrões de Hearst nos processos de organização do conhecimento na literatura brasileira da CI.

1.1 DELIMITAÇÃO DO PROBLEMA E TESE

A influência da crescente melhoria, com o advento da tecnologia, a favor da comunicação e recuperação, torna também progressiva a quantidade de informação para ser disseminada. Tal fenômeno influencia, por seu turno, a criação de novos procedimentos e instrumentos para melhorar a qualidade de publicação, de modo que a criação do processo editorial é lançada para suprir esta necessidade. Mesmo nessas condições de avaliação por pares dos periódicos científicos eletrônicos, a quantidade de publicações vem crescendo aceleradamente, levando-se em consideração que o mesmo acontece com as outras formas de comunicação científica, como os anais de eventos, o que torna ainda mais imprecisos os resultados de recuperação de informação.

No atual momento dos periódicos científicos eletrônicos, com todas as dinâmicas proporcionadas no contexto da *Web*, mesmo utilizando diversos indexadores, em muitos casos, a recuperação de informação não é precisa, considerando-se a proposta de precisão do modelo booleano. Isto é decorrente de uma série de fatores que envolvem parte do processo de recuperação, representação do documento do *corpus* e expressão de busca, uma vez que os documentos não são necessariamente representados de maneira similar às futuras propostas de expressão de buscas. Por outro lado, o usuário não conhece a forma de contextualizar a sua expressão de busca. Nesse contexto, é ressaltada a dificuldade de representar por meio de uma lógica matemática o processo de recuperação da informação.

O grande desafio para ambas as ciências que trabalham interdisciplinarmente com a recuperação de informação, relaciona-se com as perspectivas de busca e recuperação em periódicos científicos eletrônicos por meio dos SOC. Cabe considerar que a "gigante da *Web*" Google já tem indícios de aperfeiçoamentos no processo de indexação de periódicos científicos eletrônicos e ainda aumenta grandemente a chance de um artigo científico ser relevante para a necessidade de informação do usuário.

Esta pesquisa tem como área a CI, observando-se o domínio da organização e representação do conhecimento e o subdomínio dos sistemas de organização do conhecimento com foco em vocabulários controlados aplicados a periódicos científicos eletrônicos. As questões que norteiam o estudo surgiram a partir de observações, na literatura, entre os tipos de documentos: teses e dissertações, artigos científicos, livros, documentação de *softwares* e sistemas de interoperabilidade. Para efeitos deste estudo, foi realizada uma coleta de dados em teses e dissertações, na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), como forma de orientar o delineamento da pesquisa. Diante da falta de pesquisas científicas,

evidencia-se o interesse no desenvolvimento da temática, considerada de relevância, tendo como suporte a pesquisa de Santos (2015) para subsidiar a proposta teórico-metodológica deste estudo.

A escolha pela BDTD justifica-se pela relevância e representatividade dessa base nacional das pesquisas desenvolvidas em nível *stricto sensu*. Como suporte ao delineamento do estudo, realizou-se, nessa base, uma pesquisa com interesse em relacionar trabalhos correlatos no que tange a interoperabilidade de SOC. A estratégia de busca adotada foi: *("interoperabilidade" OU "mapeamento" OU "compatibilidade" OU "conversibilidade" OU "intercâmbio") E ("sistemas de organização do conhecimento" OU "vocabulário controlado" OU "tesauro" OU "ontologia")*, de que resultaram 14 registros com os termos de busca no título. Destes, em análise do título e resumo, somente três trabalhos estavam relacionados com a interoperabilidade de ontologias ligada à área da computação (FELICÍSSIMO, 2004; HINZ, 2008; DALTIO, 2007), e somente um trabalhou a temática de mapeamento e interoperabilidade em SOC na área da CI (ANDRADE, 2015).

Felicíssimo (2004) propôs em seu trabalho uma estratégia de interoperabilidade entre ontologias com a utilização do Componente para Alinhamento Taxonômico de Ontologias (CATO), que realiza de maneira automática o alinhamento das ontologias. Esse alinhamento acontece por meio de três etapas: a) comparação lexical dos conceitos com mecanismo de poda estrutural como parada; b) comparação das estruturas hierárquicas das ontologias; c) refinamento dos resultados das comparações por meio da categorização "similares", "bem similares" ou "pouco similares".

Daltio (2007) trata de um serviço *Web* (*Webservice*) para ontologias denominado Aondê[2], que suporta várias operações com a finalidade de proporcionar a interoperabilidade de ontologias distribuídas. O protótipo foi proposto para dar suporte nas consultas realizadas ao *WeBios* - sistema de informação de Biodiversidade da IC-Unicamp. A integração de ontologias é um dos processos importantes na definição de relações e representações heterogêneas. Entre as abordagens estão: mapeamento, união e alinhamento, Kalfoglou e Schorlemmer (2003) e Bruijn, Martin-Recuerda, Manov e Ehrig (2004) citados por Daltio (2007). Daltio (2007) ainda aborda a identificação de similaridades de ontologias, com a finalidade de verificar a ligação de partes ontológicas (equivalências e relações). Rahm e Bernstein (2001) são citados por Daltio (2007) para classificar as duas técnicas de similaridade: Técnicas de combinação em nível de elemento e Técnicas em nível de estrutura.

[2] Segundo Daltio (2007), Aondê significa "coruja" em Tupi, que é uma referência à linguagem de representação de ontologias em OWL.

Não foi possível localizar a dissertação de Hinz (2008) em função da indisponibilidade do material. Em busca a outras bases, contudo, foi possível recuperar o artigo publicado no Ontobras 2008 de autoria de Hinz e Pallazo (2008), o qual faz uma descrição resumida *a priori* de como ocorre a interoperabilidade dinâmica entre ontologias, por meio de algoritmos com operadores relacionais de união, intersecção, diferença e relacionado-com[3]. Exemplifica em um ambiente genérico, com uma ontologia em que seus dados são descritos por metadados *Dublin Core* e um agente com sua ontologia descrita em *FOAF (Friend Of A Friend).*

Andrade (2015) aborda a questão do mapeamento entre SOC de termos, conceitos e equivalências para construção de estratégias de busca em bases de dados, por meio de SOC interoperáveis. Sua problemática está focalizada nos relatos de SOC mapeados e interoperáveis, formato de organização de resultados originados das estratégias de buscas e sua disponibilização aos pesquisadores. A realização do estudo de caso se dá na área da Saúde e subárea Ortopedia e Traumatologia, a partir do mapeamento dos termos dos instrumentos Descritores em Ciência da Saúde (DeCS), o *Medical Subject Headings* (MeSH) do PubMed e o tesauro da *Library and Informations Science Abstracts* (LISTA).

No Quadro 1 apresenta-se a relação dos trabalhos relacionados com a temática.

Quadro 1 - Síntese dos trabalhos nacionais correlatos

Autor	Título	Nível	Instituição	Ano
Carolina Howard Felicíssimo	Interoperabilidade semântica na Web: uma estratégia para o alinhamento taxonômico de ontologias	Dissertação	PUC-RIO	2004
Jaudete Daltio	Aondê: um serviço Web de ontologias para interoperabilidade em sistemas de biodiversidade	Dissertação	UNICAMP	2007
Verlani Timm Hinz	Algoritmos para Interoperabilidade entre Ontologias	Dissertação	UCPEL	2008
Julietti de Andrade	Interoperabilidade e mapeamentos entre sistemas de organização do conhecimento na busca e recuperação de informações em saúde: estudo de caso em ortopedia e traumatologia	Tese	USP	2015

Fonte: Dados da pesquisa.

Esta ambientação dos trabalhos nacionais correlatos (Quadro 1) ao estudo proposto, bem como os relatos são importantes para a observação e delineação da pesquisa, devido ao estudo estar relacionado com os vocabulários em língua portuguesa. Considera-se que no âmbito da ontologia existem avanços tecnológicos para mapeamento de interfaces entre

[3] Nome do operador relacional de ontologias atribuído por Hinz e Pallazo (2008).

ontologias, bem como ferramentas de integração, alinhamento, combinações de ontologias. Por outro lado, percebe-se a inexistência de estudos de interoperabilidade nos demais instrumentos de SOC, como tesauro, taxonomia, esquemas de classificação, lista de cabeçalhos de assuntos e listas de termos autorizados. A única pesquisa recuperada que aborda o SOC, exceto a ontologia, é da autora Andrade (2015), na qual ainda se percebem processos manuais ou semiautomáticos para mapeamento.

Diante deste contexto, surge a questão que se buscou responder com os resultados: como recuperar a informação em periódicos científicos eletrônicos por meio de vocabulários controlados mapeados a partir das palavras-chave em um modelo de interoperabilidade semântica? A interoperabilidade é caracterizada por vários níveis e/ou tipos: Andrade e Cervantes (2012) citam a técnica, semântica, organizacional, política e humana, intercomunitária, legal e internacional; Zeng (2019) cita a sistêmica, semântica, sintática e estrutural; Moreiro González *et. al* (2012), técnica, semântica e organizacional. A opção deste estudo pela interoperabilidade semântica de SOC foi de proporcionar modelo de interoperabilidade para organização e representação do conhecimento e uma interface para o usuário, por meio de um vocabulário comum, efetuar a sua busca.

Por conseguinte, a *hipótese* é de que os vocabulários controlados aplicados a periódicos científicos eletrônicos gerenciados pelo OJS podem ser considerados como um instrumento interoperável para melhorar os processos de busca e recuperação dos artigos, e igualmente ser base para os autores consultarem os melhores termos que representam sua pesquisa. Diante deste contexto, *defende-se a tese* de que a necessidade de controle de vocabulário interoperável num contexto contemporâneo de grande volume de produções científicas é de extrema urgência e torna-se indispensável o acompanhamento deste crescimento científico, juntamente com a produção de instrumentos para tratamento e aperfeiçoamento da recuperação da informação.

1.2 JUSTIFICATIVA

Diante do objetivo geral deste estudo, entendem-se como necessários e de extrema importância os aspectos de interoperabilidade entre os vocabulários controlados dos periódicos da área da Comunicação e Informação. Existe uma tendência de pesquisa relacionada com a interoperabilidade de vocabulários controlados com vistas à disponibilização para o acesso aberto e dados ligados, e que ainda correlaciona estes acontecimentos com o *big data*. Este emaranhado de tecnologias e sistemas da informação torna-se desafiador na atuação dos

profissionais das áreas relacionadas com o desenvolvimento de recursos para organização do conhecimento neste cenário de globalização de dados e sua massificação.

Este estudo justifica-se pela importância de criar propostas de melhorias na linguagem de representação e ao processo de organização e representação do conhecimento, em especial no que tange aos periódicos científicos eletrônicos, especificamente as palavras-chave, mantendo o foco centrado no usuário dos referidos periódicos. O mesmo se aplica no sentido de produzir a motivação dos autores em representar, da melhor forma, o conteúdo do seu artigo por meio do vocabulário controlado do periódico. Outro fator de incentivo deste estudo está na continuidade da pesquisa iniciada no Programa de Pós-Graduação em Ciência da Informação – Mestrado Acadêmico da Universidade Estadual de Londrina (UEL), a qual possibilitou a construção da ferramenta e sua aplicação de maneira experimental na revista Informação & Informação, vinculada ao próprio Programa de Pós-Graduação acima citado.

Por outro lado, o vínculo estabelecido entre duas áreas de conhecimento interdisciplinares, Tecnologia da Informação (TI) e CI, instiga a produção de contribuições relevantes às áreas aplicadas. Outra motivação está nas áreas em que poderão ser aplicados os resultados desta pesquisa. Vale ressaltar a projeção de desenvolvimento para a área da Organização e Representação do Conhecimento (ORC), com as contribuições das discussões e aporte da consolidação de uma proposta de organização e visualização das palavras-chave para a recuperação dos artigos por meio do vocabulário controlado do periódico.

Desse modo, esta pesquisa pretende contribuir, de maneira relevante, para a área da Comunicação e Informação com o desenvolvimento de um protótipo de instrumento de organização do conhecimento interoperável. Não obstante, o método científico aplicado poderá servir de suporte para implantação em outras áreas do conhecimento. A comunidade científica que integra os processos de produção científica poderá obter benefícios na instanciação da submissão do seu artigo científico, devido às orientações e ao instrumento para representar, por meio da atribuição de palavras-chave significantes aos artigos científicos. No sentido oposto da produção científica, encontra-se o leitor, que executa os processos de busca e recuperação de artigos científicos.

A epistemologia da área do conhecimento em questão, por meio deste estudo, poderá obter contribuições, como a organização e perfil de pesquisa da área e subárea do conhecimento. O instrumento proposto é construído conforme os procedimentos metodológicos, a partir de termos extraídos de palavras-chave dos artigos científicos da área, e validado com um instrumento de controle de vocabulário de uma das subáreas - CI - e da outra subárea, Comunicação, desenvolvido a partir da compilação dos termos das palavras-chave dos artigos.

Esta organização do conhecimento está vinculada com a produção do conhecimento científico. Soma-se à pertinência da pesquisa, a organização e representação do conhecimento científico para fins de busca e recuperação de maneira eficiente e eficaz, a partir de vínculos estabelecidos com os conceitos rotulados em instrumento de controle de vocabulário.

Para a comunidade científica e a comunidade externa, o mapeamento científico da área, *a priori,* é de extrema importância. Por outro lado, este instrumento possibilita externar o perfil das pesquisas que contribuem para o fortalecimento da área. A pesquisa sobre a compatibilização das palavras-chave com os vocabulários controlados, integrada com as análises dos resultados, poderá promover surpreendentes agrupamentos, mapeamentos, com variantes diversificadas dos periódicos científicos eletrônicos. Poderá gerar uma base de dados de conhecimento dos periódicos científicos gerenciados pelo sistema OJS, com vistas ao aperfeiçoamento do processo de organização e representação para a recuperação dos artigos científicos por meio de um vocabulário controlado. Para essa finalidade, procura-se demonstrar o perfil dos periódicos contemplados pelo estudo.

Cabe situar esta pesquisa também no contexto da agenda de 2030 das Nações Unidas, que publica 17 objetivos do desenvolvimento sustentável, incluindo o desenvolvimento econômico, ambiental e social. Por meio da tecnologia de informação e acesso ao conhecimento, visa promover o desenvolvimento sustentável e amelhora da qualidade de vida das pessoas. Parte-se do pressuposto de que a ORC pode contribuir significativamente para os Objetivos de Desenvolvimento Sustentável (ODS). A *International Federation of Library Associations and Institutions* (IFLA) publicou, em seu documento intitulado Acesso e Oportunidade para Todos - Como as bibliotecas contribuem para a agenda de 2030 das Nações Unidas, traduzido pela Federação Brasileira de Associações de Bibliotecários e Instituições (FEBAB), a relação dos ODS com a integração dos serviços prestados pelas unidades de informação. Portanto, a proposta de organização e representação do conhecimento deste estudo, visando atender a necessidade de informação, alinha-se com os ODS (IFLA, 2015).

A demanda por estudos de interoperabilidade vai além das questões semânticas e sintáticas, o que são possíveis de serem operacionalizadas a partir de um *software* gestor de vocabulários controlados, estudos dessa natureza, contudo, ainda são escassos. A interoperabilidade entre interface de busca e o instrumento instanciado no sistema de gestão de vocabulários controlados, torna-se uma importante contribuição para a área de conhecimento. Esta contribuição é beneficiada pela organização e representação da informação com proposição na busca e recuperação mais precisa. Vale ressaltar essa necessidade de um modelo de interoperabilidade voltado para o uso dos periódicos científicos eletrônicos gerenciados pelo

OJS, que tem implantado a VCPC *Tools* como recurso de controle de vocabulário a partir das palavras-chave atribuídas pelos autores aos artigos.

Por outro lado, justifica-se a contribuição deste estudo quando integrado à finalidade dos grupos de trabalho e pesquisa: Linguagem Unesp, Georc (Unesp) e Vocabulários controlados em periódicos científicos eletrônicos (UEL), nos quais se discute a construção de SOC e suas aplicações com a perspectiva de interoperabilidade. O grupo de trabalho Linguagem Unesp tem como atividade o desenvolvimento e manutenção do Tesauro Unesp e integra, nos planos de atividades, as ações relacionadas com o desenvolvimento de manuais, compatibilizações de macroestruturas. No grupo de pesquisa Georc, discutem-se os aspectos teóricos relacionados aos fundamentos de SOC, e o grupo de pesquisa Vocabulários controlados em periódicos científicos eletrônicos destina-se às discussões em torno da implantação e manutenção dos vocabulários controlados.

Destacam-se as contribuições dos grupos de pesquisa e de trabalho no desenvolvimento desta tese. Sistematicamente a pesquisa inicia-se com uma ambientação teórica, requisito fundamental proporcionado nas reuniões do Georc, que com as atividades teóricas possibilitou conhecer e discutir esses aspectos e incluí-los no referencial de sustentação desta tese. Por outro lado, o Grupo de Trabalho de Linguagem e o Grupo de Pesquisa em vocabulário controlado da UEL proporcionaram discussões e registros de práticas precisas no que se refere às questões metodológicas de aplicação de vocabulários controlados, cada grupo em seu ambiente aplicado. Com estas contribuições, a modelagem da proposta teórico-metodológica tomou consistência e formato.

1.3 OBJETIVOS

O objetivo geral foi apresentar uma proposta teórico-metodológica destinada a subsidiar os processos de implantação e de gestão da interoperabilidade semântica entre vocabulários controlados aplicados em periódicos científicos eletrônicos.

Os objetivos específicos foram:

- discutir os aspectos relacionados aos vocabulários controlados enquanto sistemas de organização do conhecimento e sua aplicação na gestão de palavras-chave em periódicos científicos eletrônicos;

- explorar a utilização de padrões léxico-sintáticos como subsídios para mapeamento e compatibilização de palavras-chave com a sistemática do vocabulário controlado;
- delimitar o conceito de interoperabilidade no âmbito de sua aplicação em vocabulários controlados para periódicos científicos e seus aspectos relacionados ao modelo de dados e à linguagem, com foco nos processos de mapeamento e compatibilização;
- apresentar um modelo de interoperabilidade que inter-relacione semanticamente os vocabulários dos periódicos científicos eletrônicos e o Tesauro Brasileiro de Ciência da Informação (TBCI);
- aplicar o modelo de interoperabilidade entre vocabulários a um conjunto específico de periódicos (estudo de caso).

1.4 METODOLOGIA

A tese está sedimentada em duas seções de métodos e procedimentos. A primeira seção é caracterizada pela pesquisa que envolve os aspectos basilares de fundamentação da proposta teórico-metodológica, para desenvolvimento e aplicação no estudo de caso. A caracterização da pesquisa de fundamentação é de natureza descritiva com análise mista quali-quantitativa. O *corpus* é constituído por artigos de periódicos, trabalhos de anais de eventos e livros, recuperados por meio de pesquisas em bases de dados referenciais de disciplinas cursadas e em discussões em grupos de pesquisa. As discussões realizadas no âmbito dos grupos de pesquisa foram de fundamental importância para o delineamento dos aspectos teórico-metodológicos. A relevância dos autores e relatos de pesquisa que foram selecionados para compor os trabalhos nas reuniões dos grupos de pesquisa alinhou-se com a finalidade deste estudo. De modo complementar, o desenvolvimento das pesquisas metodológicas aplicadas contribuiu significativamente para os aspectos metodológicos do estudo de caso desta tese, bem como a execução das atividades e ações instanciadas.

Diante da finalidade do estudo relacionado ao processo de validação de uma proposta teórico-metodológica, adota-se o estudo de caso como recurso estratégico de aplicação desta pesquisa, pelo qual, segundo Yin (2010), se procura resolver as questões ‘como’ e ‘por que’ ligadas à característica exploratória da pesquisa. A segunda seção destina-se a descrever as etapas de modelagem desta proposta: reconhecer o ambiente de vocabulário controlado de

periódicos científicos eletrônicos – OJS; estruturar o conteúdo dos artigos do periódico científico eletrônico em formato de busca em texto completo; identificar os padrões de Hearst (1992; 1998) a partir das palavras-chave não compatibilizadas; mapear, por meio dos padrões de Hearst (1992; 1998), a sistemática das palavras-chave com o vocabulário controlado e compatibilizar com os níveis de correspondência; desenvolver o mapeamento reverso das palavras-chave compatibilizadas pelos processos de Santos (2015) a partir dos níveis de correspondência. Essas etapas foram executadas para a aplicação da proposta teórico-metodológica no portal de periódicos selecionado como *corpus* de análise, a descrição de cada uma com maior detalhamento está no capítulo 5.

1.5 SISTEMÁTICA DA TESE

A tese está organizada em três capítulos de fundamentação, um capítulo dedicado à proposta teórico-metodológica, um capítulo descrevendo a aplicação da proposta no estudo de caso e seus resultados e discussões, e ao fim um capítulo dedicado às considerações finais. O Quadro 2 apresenta a organização da tese.

Quadro 2 - Sistematização da Tese

OBJETIVO GERAL Apresentar uma proposta teórico-metodológica destinada a subsidiar os processos de implantação e de gestão da interoperabilidade semântica entre vocabulários controlados aplicados em periódicos científicos eletrônicos.	
CAPÍTULOS	**OBJETIVOS ESPECÍFICOS**
1 Introdução	-
2 Vocabulário controlado como instrumento de busca e recuperação da informação em periódico científico	Discutir os aspectos relacionados aos vocabulários controlados enquanto sistemas de organização do conhecimento e sua aplicação na gestão de palavras-chave em periódicos científicos eletrônicos.
3 Estrutura sistemática de vocabulários controlados com subsídios da linguística e da aplicação de padrões léxico-sintáticos	Explorar a utilização de padrões léxico-sintáticos como subsídios para mapeamento e compatibilização de palavras-chave com a sistemática do vocabulário controlado.
4 Modelo e linguagem de interoperabilidade entre vocabulários controlados	Delimitar o conceito de interoperabilidade no âmbito de sua aplicação em vocabulários controlados para periódicos científicos e seus aspectos relacionados ao modelo de dados e à linguagem, com foco nos processos de mapeamento e compatibilização.

Continua

Continuação

5 Proposta teórico-metodológica de interoperabilidade	Apresentar um modelo de interoperabilidade que inter-relacione semanticamente os vocabulários dos periódicos científicos eletrônicos e o Tesauro Brasileiro de Ciência da Informação (TBCI).
6 Estudo de caso Portal de periódicos da Comunicação e Informação da UEL	Aplicar o modelo de interoperabilidade entre vocabulários a um conjunto específico de periódicos (estudo de caso)
7 Considerações finais	-

Fonte: Elaborado pelo autor (2020).

O segundo capítulo aborda os aspectos históricos do percurso do periódico científico impresso ao eletrônico; os principais subsídios da recuperação da informação e seus modelos clássicos aplicados aos periódicos científicos eletrônicos de organização; o SOC e fundamentos da representação da informação; e a disponibilização dos vocabulários controlados, gerados a partir das palavras-chave dos periódicos científicos eletrônicos gerenciados pelo OJS, com base no currículo *Lattes* de autores que tratam da temática de vocabulários controlados em periódicos científicos eletrônicos.

O terceiro capítulo trata dos aspectos fundamentais de sistemática com abordagem em construção de vocabulários controlados; dos aspectos linguísticos e linguística de *corpus*; das recomendações de processos de desenvolvimento de vocabulários controlados a partir da norma ISO 25.964:1; e dos padrões de Hearst (1992, 1998) para a identificação de hierarquia de conceitos e termos.

O quarto capítulo fundamenta os modelos de dados; as linguagens de interoperabilidade entre vocabulários controlados; as características operacionais da norma ISO 25.964:2; os modelos de dados SKOS e *Webservices* e *Application Programming Interface* (API) de *softwares* gestor de vocabulários controlados. Descreve, ainda, os tipos de interoperabilidade entre SOC - vocabulários controlados.

O quinto capítulo contempla a estrutura da proposta teórico-metodológica com cinco etapas de atividades para atender aos objetivos específicos desta tese, e o sexto capítulo descreve os respectivos resultados e efetua análises referentes à aplicação do estudo de caso, por meio do qual se mostra a execução das cinco etapas previstas na proposta. No sétimo capítulo apresentam-se as considerações finais.

2 VOCABULÁRIO CONTROLADO COMO INSTRUMENTO DE BUSCA E RECUPERAÇÃO DA INFORMAÇÃO EM PERIÓDICO CIENTÍFICO

O contexto histórico do periódico científico eletrônico na comunicação científica é expresso em fatos que perpassam a massificação de artigos científicos nos fenômenos relacionados ao aumento expressivo de informações na *Web*. Os aspectos da recuperação da informação são amparados pelas teorias que tratam dos principais modelos denominados como clássicos: booleano, vetorial e probabilístico.

Esses modelos inserem-se nos mecanismos de busca e recuperação no contexto do periódico científico eletrônico gerenciados pelo OJS. Os fundamentos de SOC e o tratamento temático por meio do vocabulário controlado são identificados nos periódicos gerenciados pelo OJS. Esses periódicos utilizam o vocabulário controlado para representação do conteúdo nos artigos científicos pelos autores, bem como na interface de busca integrada ao vocabulário controlado.

2.1 PERIÓDICO CIENTÍFICO ELETRÔNICO: PRINCIPAIS ABORDAGENS

Presume-se remeter ao século XVII o marco do surgimento do periódico científico, pois, como visto, não se tem ao certo quando ocorreu, mas esse acontecimento se deu em sua segunda metade, ou seja, 1650-1700 (aproximadamente), com a iniciação da comunicação científica juntamente com a propagação da imprensa escrita. Em meio às diversas evoluções, o periódico científico passou a integrar o suporte computacional, o que originou o periódico científico eletrônico. Suas características fundamentais ultrapassam os limites do papel e integram-se a recursos como hipertexto – ainda que pouco utilizado, links e multimídias, ou seja, um formato de comunicação científica dentro de um modelo de processos eletrônicos de editoração. Os primeiros experimentos foram entre 1978 e 1980, com a *Eletronic Information Echange System* (EIES). Já no Brasil algumas iniciativas aconteceram em 1990.

O periódico científico, uma modalidade de comunicação científica, passa por diversas evoluções, assim como a própria comunicação científica. A comunicação científica e o conhecimento, não obstante ter ocorrido oralmente, desde a Grécia Antiga, já existiam documentos registrados para fins científicos. Configura-se a existência de bibliotecas públicas com livros escritos em papiro e conservados em caixas e, com a evolução, os papiros passam a ser substituídos por pergaminhos de origem animal (CÔRTES, 2006). Meadows (1999)

complementa que eram realizados os registros por meio de manuscritos dos debates científicos e para disseminá-los seria necessário, literalmente, copiar os manuscritos.

Com a propagação da imprensa, a comunicação científica também sofre uma evolução e crescimento. Está integrada com o início de pesquisas científicas disseminadas de diversas formas, principalmente de maneira verbal e escrita (MEADOWS, 1999). A evolução na tecnologia aplicada na impressão e a massificação de produção de papel contribuíram para que a comunicação escrita se tornasse um importante meio de disseminação científica. Esbarrou nas dificuldades de distribuição geográfica e número reduzido de exemplares (CÔRTES, 2006). Conforme Lara (2006) e Mueller (2000a), estes materiais impressos, como livros e periódicos, caracterizam-se comunicação formal. Mesmo diante da quantidade expressiva de impressões de comunicação científica formal, prevalece a comunicação oral em reuniões científicas, colóquios, seminários, e outro meio de disseminação, como a troca de cartas entre pesquisadores (CÔRTES, 2006), denominada como informal por Lara (2006) e Mueller (2000a).

A comunicação científica, portanto, é um conjunto de atividades voltadas a construção e comunicação do conhecimento científico, a fim de promover a evolução da humanidade (LARA, 2006; WEITZEL, 2006). A comunicação científica, para Côrtes (2006), considerando o momento do estudo, contava com uma diversidade de instigações e novas perspectivas. Nas publicações impressas, contabilizam os jornais científicos on-line, fóruns de discussão, sistemas de *open archives e open access*, entre outros.

Na comunicação científica formal inserem-se as revistas científicas, também denominadas como periódicos científicos, com grande importância para a construção do conhecimento por seu modelo ágil de disseminação. A sistemática de revisão e aprovação por pares confere a devida credibilidade ao estudo e à ciência (GONÇALVES; RAMOS; CASTRO, 2006). Surgem as primeiras revistas científicas, a *Philosophical Transactions da Royal Society*, em Londres, e o *Journal des Savants*, em Paris, no ano de 1660. O *Journal des Savants* fornecia mecanismos de divulgação de notícias e acontecimentos da República das Letras na Europa (MEADOWS, 1999; BURKE, 2003). A origem da *Royal Society* foi no fim do período de guerra de aproximadamente 20 anos, durante a restauração da monarquia em Londres, com o surgimento de pequenos grupos para debater questões relacionadas à filosofia, esquivando-se de assuntos polêmicos. Os membros da *Royal Society* faziam viagens para coleta de informações e construíam resumos da literatura publicada – este processo marca o surgimento da revista científica (MEADOWS, 1999).

Percebe-se, nos aspectos históricos das revistas científicas, o nascimento de um instrumento para tornar a comunicação científica mais rápida. Diante desta demanda, concebe-se, com o surgimento e popularização dos recursos computacionais, uma evolução no processo sistemático das revistas científicas, com a inclusão das ferramentas computacionais e de comunicação por meio das redes de telecomunicações.

Doravante, neste estudo, toma-se como padrão o termo periódico científico, tomando como base o Glossário – Termos e Conceitos da Área de Comunicação e Produção Científica de Marilda Lopes Ginez de Lara (org.), que o trata como sinônimo de revista científica (LARA, 2006). O conceito de periódico científico eletrônico é o de um documento eletrônico para publicação que se beneficia dos recursos imagéticos, áudios, vídeos, *links*, convertendo, estruturalmente, a comunicação científica de maneira tradicional em formato hipertexto. O periódico científico eletrônico depende de um suporte eletrônico, ou seja, na maioria dos casos o suporte é a internet, mas também pode ser disponibilizado em CD-ROM, entre outras mídias eletrônicas (LARA, 2006; GONÇALVES; RAMOS; CASTRO, 2006; MUELLER, 2000b).

Por mais que exista proximidade entre a definição de documento eletrônico e de documento digital, é importante mencionar as diferenças. Todo documento digital[4] é eletrônico, mas nem todo documento eletrônico[5] é digital, ou seja, o documento digital é um termo mais específico, faz parte do grupo de documentos eletrônicos (BODÊ, 2016), porém são tratados como sinônimos por Cunha e Cavalcanti (2008). São exemplos de documentos eletrônicos não digitais aqueles que dependem somente da eletrônica para codificá-lo e decodificá-lo, como filme em VHS, música em fita cassete.

Os periódicos científicos eletrônicos não têm as mesmas características e nem os mesmos formatos. Quando comparados a um meio comum de comunicação, podem ser categorizados em três grupos: formato *Portable Document Format* (PDF), semelhante ao formato impresso; formato eletrônico do texto impresso com recursos adicionais como navegação, hipertextual e multimídia; e publicação exclusivamente em meio eletrônico, mesmo aplicando pouco os recursos hipertextual e multimídia (GONÇALVES; RAMOS; CASTRO, 2006).

[4] De acordo com o Glossário da Câmara Técnica de Documentos Eletrônicos (2014, p.19), o documento digital é a "informação registrada, codificada em dígitos binários, acessível e interpretável por meio de sistema computacional".

[5] Câmara Técnica de Documentos Eletrônicos (2014, p.19) define o documento eletrônico como "informação registrada, codificada em forma analógica ou em dígitos binários, acessível e interpretável por meio de um equipamento eletrônico."

Em relação ao hipertexto, Moreira (2003) trata dos aspectos de reprodução e criação. Tanto para os documentos impressos como para aqueles que já nascem no formato digital, existem três processos básicos de transmutação: "a transcrição, a tradução e a criação, que aliás, não são subsequentes, nem mesmo complementares" (MOREIRA, 2003, p.17).

A transcrição é a mais rudimentar aos periódicos científicos eletrônicos, na grande maioria, são produzidos digitalmente, mas planejados para o suporte impresso e posteriormente convertidos para *HyperText Markup Language* (HTML), nesse processo tenta-se manter a ordem canônica do texto. Já a tradução consiste em adaptar o texto, consideram-se a linguagem natural e a linguagem computacional, basicamente utiliza-se de sua superestrutura. A criação consiste em estabelecer links, tanto os semânticos como os referenciais, portanto é a principal diferença do processo de criação de textos (MOREIRA, 2003).

Neste contexto de disseminação do periódico científico eletrônico, ressaltam-se as questões relativas à recuperação da informação nos sistemas de repositórios de artigos.

2.2 RECUPERAÇÃO DA INFORMAÇÃO E SEUS MECANISMOS EM PERIÓDICO CIENTÍFICO ELETRÔNICO

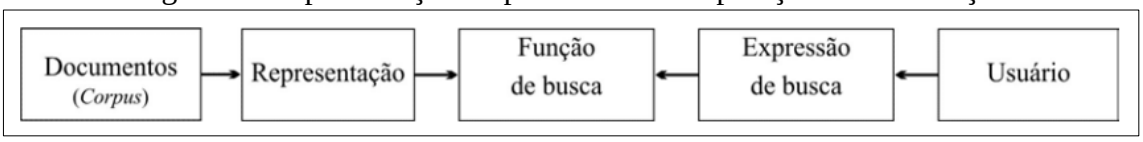
O termo “recuperação da informação” ou *information retrieval* foi instanciado a partir da pesquisa de Calvin Mooers em 1951, quando publica o seu artigo no periódico *American Documentation*, que então registra a primeira aparição do termo. Mooers (1951) utiliza a expressão “recuperação da informação” para nomear um processo ou método pelo qual o usuário expressa sua necessidade de informação para um conjunto de documentos que contenham a informação de acordo com a sua expressão de necessidade. De maneira geral, é a denominação para uma solicitação de bibliografia, também chamada de processo de descoberta de informação em algum local de armazenamento. A recuperação de informação tem abrangência nos aspectos intelectuais da descrição de informação e nas especificações para uma pesquisa, incluem-se, nesse âmbito, os sistemas e máquinas empregadas para operacionalizar esta atividade.

Ferneda (2003, p.14) aborda que a recuperação da informação tem vários significados na CI, entre os quais estão: “[...] a operação pela qual se seleciona documento, a partir do acervo, em função da demanda do usuário”; “[...] fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes [...]”. Em comum, essas definições concentram-se no estudo do processo de busca de informação. Esse processo identifica, num dado conjunto de elementos, aqueles que atendem à necessidade de informação

do usuário. Os modelos clássicos relacionados à recuperação da informação, como já mencionado, são: booleano, vetorial e probabilístico.

O processo de recuperação de informação consiste em "[...] identificar, no conjunto de documentos (*corpus*) de um sistema, quais atendem à necessidade de informação do usuário" Ferneda e Dias (2013, p. 53). Na Figura 2, apresenta-se uma representação simplificada desse processo. A função de busca é centralizada no processo e tem como finalidade compatibilizar a representação com a expressão de busca.

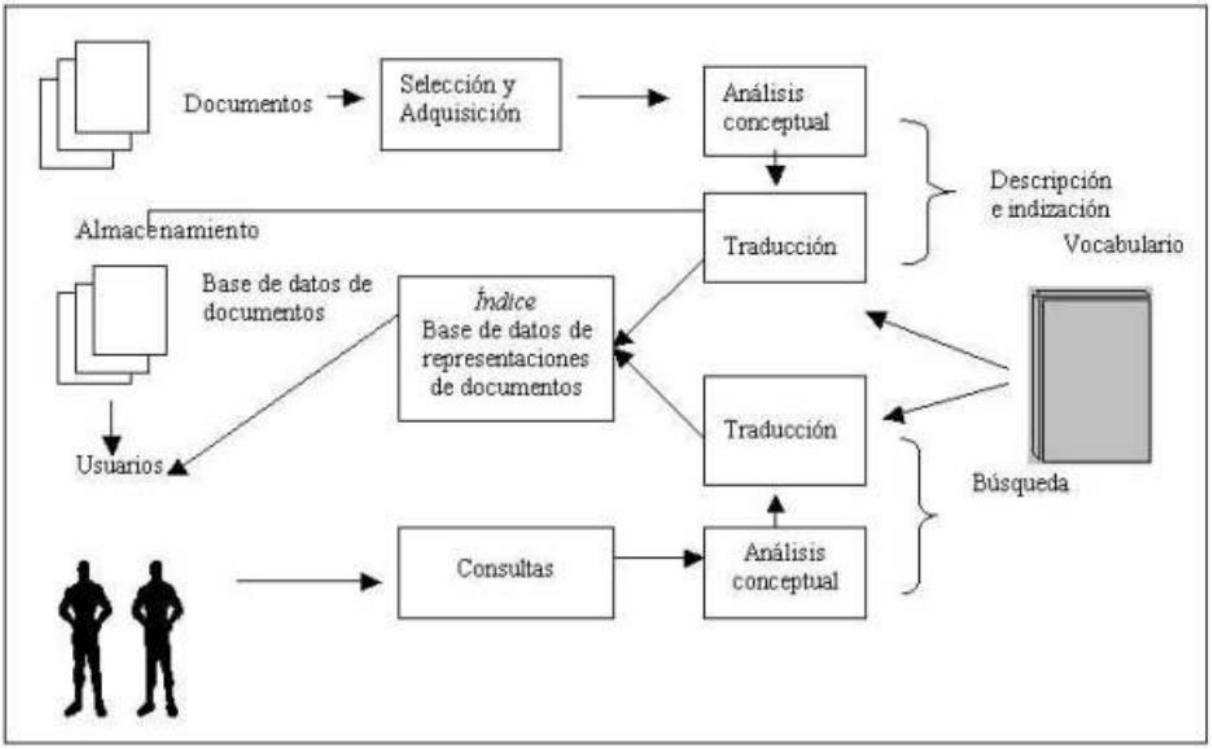
Figura 2 - Representação do processo de recuperação de informação



Fonte: Ferneda (2003, p.15)

A Figura 3 representa os principais componentes de um sistema de recuperação da informação de Lancaster (1995) que, neste sentido, corrobora com a representação simplificada de Ferneda (2003).

Figura 3 - Principais componentes de um sistema de recuperação da informação



Fonte: Lancaster (1995, p.18)

Lancaster (1995) detalha os processos implícitos na representação da informação e inclui o vocabulário controlado como instrumento para representar os documentos. Percebe-se que são representadas uma saída e uma entrada, o profissional da informação representa o

documento com uma lista de termos controlados a partir de conhecimentos do assunto tratado no documento, bem como as estratégias de buscas que os usuários poderão construir para recuperar o documento.

Em ambiente *Web*, acontecem mudanças em relação ao suporte do documento, sendo a principal delas a "[...] desterritorialização do documento e a sua desvinculação de uma forma física tradicional como o papel, possibilitando uma integração entre diferentes suportes (texto, imagem, som) [...]". Outra questão importante na *Web* é a condição de acesso do documento com o hipertexto, que por sua vez tem sua linearidade alterada para o acesso (FERNEDA, 2003, p.16). Dias (1999) corrobora com referido autor afirmando que a característica essencial do hipertexto é a não linearidade, e que esta característica surge em momento anterior à *Web*, com a estruturação do documento com sumários, referências bibliográficas, notas de rodapé. A tecnologia auxiliou na velocidade de acesso não linear. Esses itens classificatórios apresentados de maneira lógica levam o leitor a acessar o conteúdo de modo seletivo, em momentos atuais não são perceptíveis a sua importância, ao contrário dos manuscritos em que esta forma de interação não existia, razão pela qual a tendência é da não linearidade (DIAS, 1999).

Diante desse universo de informação, "[...] já se concebe a convergência do impresso para o eletrônico, bem como o documento nato-digital (que já surgiu no formato digital). As bases *full text* surgem como resposta à necessidade de um perfil de leitor ávido pela informação mais atual e completa possível." (DZIEKANIAK, 2010, p.46). Ferneda (2003, p. 17) corrobora contextualizando o crescimento dos chamados sistemas de texto completo ou integral e a capacidade de armazenamento que os computadores atuais têm. Nesse aspecto, a representação do documento poderia ser por meio do próprio conteúdo do texto, já que está na íntegra e não tem uma representação. Porém, juntamente com o aumento de capacidade computacional, ocorre o aumento exponencial de documentos, o que obriga a retomada da necessidade de representação dos documentos.

Por outro lado, ainda no processo de recuperação (FERNEDA, 2003), tem-se o usuário, que necessita de informação, e para satisfazer sua necessidade de informação deverá convertê-la em uma expressão de busca. O autor supracitado diz que esta expressão pode ser representada em linguagem natural ou por meio de uma linguagem artificial. A partir da execução da recuperação, deverão resultar os documentos pertinentes à sua necessidade, e caberá ao usuário refinar essa busca, considerando os resultados que são relevantes para sua necessidade informacional. Existe uma dificuldade identificada no usuário, que é a elaboração desta expressão de busca. Sabe-se que em muitos casos o usuário não participa do processo de

representação dos documentos, ou seja, ele não tem acesso às expressões ou descrições dos documentos que poderá considerar relevantes, após a execução da sua expressão de busca.

Lancaster (1995) reforça que no processo de representação é importante realizar duas etapas: análise conceitual e tradução a partir da linguagem artificial - nominada como vocabulário. Esse vocabulário pode ser uma lista de cabeçalho, esquema de classificação, um tesauro ou somente uma lista de frases ou palavras-chave autorizadas.

A atividade central do processo de recuperação de informação (FERNEDA, 2003) está na função de busca, por meio da qual acontece a compatibilização da expressão de busca, elaborada pelo usuário, e a representação do documento, elaborada automaticamente ou de maneira manual. A partir desta compatibilização, retornam os documentos que são contemplados por esta comparação. Neste contexto, cabe abordar os modelos de recuperação da informação, alguns são clássicos e mais utilizados, relacionam-se como a eficiência e eficácia dos resultados da recuperação. Os modelos agrupados como clássicos são: booleano, vetorial e probabilístico.

Os modelos quantitativos são fundamentados em cálculos matemáticos, como lógicas, estatísticas e teoria de conjuntos, portanto, para estes modelos, o documento é indexado por um conjunto de termos que o representam. Esses termos são palavras representativas de conceitos que o documento aborda, podem ser conceitos periféricos ou conceitos mais centrais do documento. Um conjunto de documentos que tem milhares de palavras que se repetem em todos eles não significa que aquele termo seja o melhor para representá-los. Por outro lado, um termo que aparece em poucos documentos pode ser um termo fortemente representativo, evita o trabalho do usuário no refinamento de muitos documentos como resultado da expressão de busca (FERNEDA, 2003). Cabe observar que neste último exemplo a necessidade do usuário pode não ser suprida. Portanto esta é uma das complexas tarefas da indexação para recuperação de informação.

O modelo booleano é “um dos modelos mais utilizados, na recuperação de informação [...]”, ele “[...] baseia-se na teoria dos conjuntos e na álgebra de Boole”. São utilizados para elaboração das expressões de buscas os operadores lógicos *and*, *or* ou *not* (KURAMOTO, 2002, p.1). Silva, Santos e Ferneda (2013, p.29) complementam que “em um sistema booleano, o conteúdo informacional dos documentos é representado por um conjunto de termos de indexação”. Logo os documentos resultantes atendem às restrições lógicas constituídas pela expressão de busca.

Outros modelos foram desenvolvidos para melhorar a recuperação de informação, como quantitativos: fuzzy e booleano estendido, e os dinâmicos: sistemas especialistas, redes

neurais e algoritmos genéticos. Porém, para efeito deste estudo, optou-se por restringir-se ao modelo booleano e ao vetorial, por se tratar dos mais próximos utilizados nos portais de busca de periódicos científicos eletrônicos, e ao modelo de análise de referência e probabilístico, que são aplicados por mecanismos de busca e têm influência direta na busca por artigos científicos.

No sistema OJS, é marcante a presença do modelo booleano para formulação de expressão de busca. Nos portais de periódicos científicos eletrônicos constam as dicas de pesquisa, como: o sistema de busca não diferencia maiúsculas ou minúsculas; termos irrelevantes são ignorados pelo sistema de busca; são recuperados por padrão apenas artigos contendo todos os termos de busca (ex.: *AND* é implícito); a combinação de múltiplos termos com *O*R tem como finalidade encontrar os artigos contendo um ou outro termo, ex.: educação *OR* pesquisa; na utilização de parênteses criam-se buscas mais complexas, ex.: arquivo ((revista *OR* conferência) *NOT* teses); no uso de aspas duplas, recupera-se o termo exato, ex.: "Acesso Livre à informação"; na necessidade de excluir um termo, utiliza-se - ou *NOT*, ex.: “online – políticas” ou “online *NOT* políticas”; por fim, tem-se o * como caractere coringa, ex.: soci*, o qual recuperará documentos contendo termos como "sociedade" ou "sociológico".

Uma expressão conjuntiva de enunciado *t1 AND t2* recuperará documentos indexados por ambos os termos (t1 e t2). Uma expressão disjuntiva *t1 OR t2* recuperará o conjunto dos documentos indexados pelo termo t1 ou pelo termo t2. Uma expressão que utiliza apenas um termo t1 terá como resultado o conjunto de documentos indexados por esse termo. A expressão *NOT t1* recuperará os documentos que não são indexados pelo termo t1. As expressões *t1 NOT t2* ou *t1 AND NOT t2* terão como resultado o conjunto dos documentos que são indexados por t1 e que não são indexados por t2. Termos e operadores booleanos podem ser combinados para especificar buscas mais detalhadas ou restritivas. Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem delimitando partes da expressão por meio de parênteses. A definição de expressões complexas exige um conhecimento profundo da lógica booleana. O conhecimento da lógica booleana é importante também para entender e avaliar os resultados obtidos em uma busca. (FERNEDA; DIAS, 2013, p.56)

Kuramoto (2002, p.2) faz uma correlação do modelo booleano com o modelo vetorial, o primeiro “[...] baseia-se na comparação exata entre os termos de uma consulta e aqueles presentes nos documentos [...]”, o segundo “[...] baseia-se na comparação parcial entre a representação dos documentos e a da consulta do usuário”. Esta comparação parcial é possível em virtude da “[...] atribuição de pesos aos termos de indexação presentes na consulta e aqueles presentes nos documentos”. Portanto este processo ocorre por meio da ponderação que cada

termo de indexação recebe e consequentemente do cálculo de similaridade. Os documentos que atendem a uma consulta têm um grau de similaridade atribuído, que impactará na apresentação dos resultados ordenados de maneira decrescente. Por outro lado, infere-se a relevância dos documentos, já que os pesos são atribuídos de acordo com a frequência dos termos da expressão de busca no documento.

Oliveira (2010) e Ferneda e Dias (2013) descrevem que, no modelo vetorial, é instanciado um vetor onde é inserido o peso ou grau de relevância de cada termo indexado para o documento no momento de representação, da mesma forma ocorre com os termos da expressão de busca. O tamanho do vetor está relacionado com a quantidade de termos constante na indexação do documento e na expressão de busca.

O modelo booleano difere do vetorial devido à forma de atribuição dos pesos e consequentemente o estabelecimento dos resultados. No booleano, os pesos são 0 e 1, ou seja, sistema binário, ao passo que no vetorial há variações entre o 0 e 1, desta forma não identifica somente a presença do termo. Zero significa não ter nenhuma relevância e 1 (um) contém total relevância, o que endossa a possibilidade de uma comparação, uma vez que os pesos não são binários (KURAMOTO, 2002).

O modelo vetorial não é marcante no mecanismo de busca nativo do OJS, pois este não está explícito, como o modelo booleano, por meio de expressões na própria orientação do sistema. Seus resultados não são fáceis de avaliar *a priori*, diante de questões intrínsecas ao modelo. De maneira sumária, pode-se comparar o modelo vetorial com a possibilidade de representar os artigos do repositório do OJS por meio de termos e realizar a busca no portal OJS utilizando o critério de <termos indexados>. Vale observar que não são explícitas, nas orientações, as informações quando se utiliza esta opção de busca e recuperação. Por outro lado, os modelos probabilístico e análise de referências não constam no rol de possíveis buscas no portal OJS, mas têm influência muito característica quando se realiza a busca por meio de um buscador comum na *Web*.

O modelo probabilístico, segundo Oliveira (2010, p. 45), "[...] trabalha com conceitos provenientes da área de probabilidade e estatística", complementa que o princípio da ordenação probabilística é a base deste modelo. Ferneda e Dias (2013, p. 58) apontam que este modelo "[...] foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores, tais como Robertson e Jones (1976)". Diante deste contexto, trata-se o processo de recuperação de informação por um processo probabilístico, devido às incertezas que o processo de recuperação de informação tem em seu entorno, portanto,

diferentemente dos modelos booleano e vetoriais, que tentam ser exatos na relevância, este pensa numa probabilidade de relevância.

O modelo probabilístico executa cálculos a partir da expressão de busca do usuário para os documentos do *corpus*. Esse resultado numérico (similaridade) representa o fator de relevância do documento, por meio dele são organizados os resultados da busca. Diante dos resultados, "[...] um primeiro conjunto de documentos, o usuário pode marcar alguns deles que considera verdadeiramente relevantes para a sua necessidade". Essa seleção de documentos poderá retornar ao sistema como relevante, e este é o conceito de *relevance feedback*. Trata-se de um fator positivo deste modelo, o único que, de maneira explícita, trabalha com o retorno do usuário para melhorias de resultados futuros (FERNEDA; DIAS, 2013, p. 58).

O protocolo de interoperabilidade é uma contribuição não somente para a troca de informações entre sistemas, mas também fator de replicação na *Web* de artigos científicos, entre outros tipos de documentos científicos, um exemplo clássico é a base de dados SciELO. Foi realizada uma busca no Google com o título dos artigos: Ferneda e Dias (2013), Kuramoto (2002) e Silva, Santos e Ferneda (2013). Os resultados das três pesquisas são muito diferentes, considerando a posição de relevância com que o artigo aparece nos resultados com remissão ao portal do periódico. No caso de Ferneda e Dias (2013), os dois primeiros resultados são de repositórios da Unesp, o endereçamento para o artigo no periódico científico aparece somente na terceira posição. Kuramoto (2002), como é um artigo de uma revista descontinuada, entre os três resultados relevantes, somente o primeiro é endereçado para o artigo, porém por meio da base de dados Brapci. No caso do artigo de Silva, Santos e Ferneda (2013), o primeiro resultado é o endereçamento para o portal do periódico e ainda traz uma indicação do Google: "modelos de recuperação de ....Mais resultados em uel.br".

A disseminação na *Web* deixa os periódicos científicos eletrônicos em evidência nos buscadores. Como referendado no modelo probabilístico, o conceito de *relevance feedback* está presente no buscador utilizado neste experimento. Torna-se implícito em seus controles de atividades do usuário para utilização de dados de buscas realizadas como mecanismo de melhorar a sua experiência e proporcionar resultados mais relevantes. Portanto não é necessário, para o usuário, marcar o melhor resultado de sua busca, entende-se o melhor resultado a partir de variantes coletadas por meio da interação com os respectivos resultados. Por exemplo, visitas frequentes, exploração imediata, entre outras relações que podem ser inferidas, fundamentadas no modelo.

Diferentemente do modelo probabilístico, o modelo que utiliza a análise de referência, abordado por Kuramoto (2002, p.2), "[...] vem sendo utilizado com sucesso por alguns

mecanismos de busca (*search engines*). O Google melhorou consideravelmente a precisão dos seus resultados utilizando esse tipo de análise.”. Portanto torna-se mais simples a manipulação dessas referências em páginas da *Web*. Em análise das duas revistas ativas, nas quais os artigos Silva, Santos e Ferneda (2013) – Informação & Informação - e Ferneda e Dias (2013) – InterScientia - foram publicados, existe um diferencial entre elas além da indexação nas diversas bases. Na Informação & Informação consta na página a inclusão do índice de citações dos artigos no Google Acadêmico. Considera-se fato a utilização do modelo de análise de citações para recuperação de informação nos portais OJS, por meio da interface do buscador.

Diante da importância de apresentar os modelos quantitativos booleanos em sistemas de recuperação da informação e a correlação destes com o periódico científico eletrônico, com ênfase no sistema OJS, cabe reforçar que os processos de recuperação nas diversas esferas são de extrema complexidade e distantes de resolução. Eles dependem não somente da representação do documento, ou nos sistemas de texto completo, de uma homogeneidade com a expressão de busca. Em ambos os casos a dependências é de usuários diferentes, ou de focos diferentes no caso de representação do documento que pode ser automatizada.

Neste processo de aperfeiçoamento da recuperação, o profissional da informação depara-se com o desafio de atuar no esforço de tratamento temático dos artigos em periódicos científicos eletrônicos com a perspectiva de busca por meio do SOC.

2.3 SOC: TRATAMENTO TEMÁTICO POR MEIO DE VOCABULÁRIO CONTROLADO

A primeira representação da iniciativa ocidental de categorizar o conhecimento concentra-se no esforço de Aristóteles para sistematizar a organização do conhecimento. A propósito, se não forem todos, uma quantidade relativamente importante do SOC com maior evidência tem uma dívida interminável com o conhecimento acumulado pelas classificações filosóficas. Desse modo, relacionam-se, por exemplo, a *Dewey Decimal Classification* (DDC) e a *Library of Congress Classification* (LCC) com a concepção filosófica de Bacon acerca da organização do conhecimento.

O SOC compreende os esquemas mais utilizados para a organização da informação e do conhecimento, incluindo os tesauros, os sistemas de classificação que são os sistemas tradicionais e conhecidos na comunidade, e ainda outros como ontologias e arquivos de autoridade aplicados para controle de nomes pessoais e geográficos que têm uma convivência menor com o público leigo, acarretando desconhecimento (HODGE, 2000).

O SOC tem sido discutido com frequência no que tange à organização e à representação do conhecimento e é usado como termo genérico para os conceitos específicos, que abrangem os citados sistemas de classificação, os tesauros, as taxonomias e as ontologias, dentre outros. Smiraglia (2014) instiga três reflexões importantes em relação ao SOC: a) seria possível proporcionar um núcleo comum entre os diversos tipos de SOC? b) por exemplo, um sistema de classificação poderia ser convertido em um tesauro? c) um tesauro poderia ser convertido em uma ontologia? As reflexões instanciadas parecem um tanto complexas de serem respondidas, mas tanto a teoria quanto a prática têm grande interesse. O interesse teórico advém da necessidade de se encontrar uma resposta precisa acerca das características que diferenciam e as que são complementares entre os diferentes tipos de SOC.

As pesquisas em Organização do Conhecimento (OC) são, em sua grande maioria, dedicadas aos aperfeiçoamentos de sistemas de classificação bibliográfica e aos tesauros, incluindo os suportes materiais e não materiais, ou seja, impressos, eletrônicos e digitais, que se encontram disponibilizados em ambiente local e/ou disseminados na *Web*. Essas linguagens de classificação aplicam-se na descrição dos registros, utilizando-se de metadados bibliográficos. Vale acrescentar que na maior parte dos sistemas de classificação, que são construídos e aplicados nas bibliotecas ou outras unidades de informação, a interação é designada socialmente de responsabilidade do profissional de informação. Portanto, ressalta-se que o usuário dessas unidades de informação raramente utiliza o sistema de organização do conhecimento para expressar sua necessidade de busca e, por conseguinte, executá-la. Por outro lado, desconhece a sua existência e acaba por recorrer aos buscadores da rede de internet (IBEKWE-SAN JUAN; BOWKER, 2017).

A descrição de objetos de conteúdos por meio de termos de um vocabulário controlado visa à consistência da descrição para uma recuperação mais facilitada e precisa. Esse controle de vocabulário tende a aumentar a eficiência e a eficácia dos sistemas de recuperação em diversos ambientes, principalmente na *Web*. A necessidade do controle de vocabulário é decorrente de duas características básicas da linguagem natural: 1) duplicidade de designação para um mesmo conceito – exemplo: SOC/Sistemas de Organização do Conhecimento; 2) duplicidade de conceito para uma mesma designação – exemplo: manga (camisa), manga (fruta). Portanto o vocabulário controlado tem como finalidade fornecer um mecanismo para organizar a informação, ocorre por meio da atribuição de termos selecionados de vocabulário controlado para descrever documentos e outros tipos de objetos de conteúdo (AMERICAN NATIONAL STANDARS INSTITUTE; NATIONAL INFORMATION STANDARDS ORGANIZATION, 2005).

O controle de vocabulário normaliza os nomes de entidades com a eliminação da ambiguidade, criando uma linguagem artificial a partir da linguagem natural. Normalmente o controle de vocabulário é implantado a partir de três momentos: selecionar o nome autorizado para a entidade; desambiguar o nome - torná-lo distinto; e instanciar nomes para demais variantes da entidade (SVENONIUS, 2000 p.89). A linguagem natural não é a mais indicada para organizar informação, uma vez que é carregada de variantes linguísticas, como sinonímia e homonímia (SVENONIUS, 2000 p.89).

O vocabulário normalmente empregado nos sistemas de recuperação é o vocabulário controlado, que *a priori* é um conjunto de termos, com limitações, utilizado por indexadores e usuários e organizado de duas formas complementares: sistemática e alfabética. A finalidade é evidenciar as relações, porém esta contribuição da organização dos termos não é o elemento principal do conceito de vocabulário controlado. O oposto do vocabulário controlado é a linguagem natural, na recuperação da informação, em que se utiliza qualquer palavra ou frase para descrever o assunto e sem limitações. Os sistemas de recuperação da informação podem trabalhar sem nenhum controle, porém algumas inconsistências nos resultados podem vir a não corresponder à necessidade de informação do usuário (LANCASTER, 1987).

2.4 VOCABULÁRIO CONTROLADO EM PERIÓDICOS CIENTÍFICOS ELETRÔNICOS: ABORDAGEM EMPÍRICA

O controle de vocabulário da revista Informação & Informação teve seus estudos fundamentais publicados por Santos e Cervantes (2014). O recorte apresentado é a análise do comportamento das palavras-chave que foram observadas, cuja coleta foi executada a partir da utilização dos artigos e de forma manual. O foco do estudo está nos processos de indexação de artigos científicos, tomando como fundamental a representação do conteúdo atribuída pelo autor, aquele que denomina as palavras-chave, sem deixar de dar a devida importância para o controle de vocabulário para a busca e recuperação de artigos no repositório eletrônico do periódico. Os resultados buscam a reflexão sobre a redundância de palavras-chave, cujas sintaxes são atribuídas no plural e também no singular. A representação temática do conteúdo do periódico é de extrema necessidade, sua eficiência é sentenciada nos resultados de busca (SANTOS; CERVANTES, 2014).

Os processos de compatibilização de palavras-chave com um vocabulário controlado com a finalidade de exercer um controle das palavras-chave utilizadas na revista Informação & Informação são executados a partir das ações: idênticas, idênticas retirando as palavras vazias, idênticas retirando os plurais, idênticas retirando os sufixos, idênticas por meio do índice contido. A remoção dos plurais e sufixos é realizada a partir de Orengo, Buriol e Coelho (2007) e de Orengo e Huyck (2001). Os resultados, após a aplicação dos processos descritos acima, são índice de termos e o perfil do periódico para os usuários finais do OJS. Reitera-se que esses produtos são resultantes da compatibilização das palavras-chave (SANTOS; CERVANTES, 2015a).

Santos e Cervantes (2015b) complementam os estudos dos autores supracitados com os resultados dos procedimentos teóricos e metodológicos para o desenvolvimento do controle do vocabulário de periódicos científicos eletrônicos, a partir da compatibilização de palavras-chave dos artigos científicos eletrônicos. A caracterização da pesquisa é de natureza descritiva com abordagem mista qualitativa e quantitativa. A coleta dos dados e a análise são realizadas de maneira automática por meio da aplicação da ferramenta VCPC *Tools*. A VCPC *Tools* foi desenvolvida para o controle de vocabulário em periódicos científicos eletrônicos de áreas especializadas e encontra-se implantada nas revistas da Ciência da Informação e da Comunicação da Universidade Estadual de Londrina. As etapas e ações do estudo são apresentadas no Quadro 3.

Quadro 3 - Sistematização da proposta de implantação de vocabulário controlado em periódicos Santos e Cervantes (2015b)

ETAPAS	**AÇÕES**
Desenvolver a ferramenta VCPC *Tools*	a) Identificar o endereço do repositório digital do periódico. b) desenvolver o projeto da ferramenta VCPC *Tools*. c) desenvolver o projeto físico do banco de dados. d) criar a base de dados no sistema gerenciador de banco de dados.
Coletar e armazenar os vocabulários controlados	a) identificar os vocabulários controlados da área especializada. b) estruturar os dados dos termos dos vocabulários controlados. c) descrever, em linguagem SQL, os termos dos vocabulários controlados. d) importar os dados dos termos dos vocabulários controlados.

Continua

Continuação

Compatibilizar as palavras-chave com os termos dos vocabulários controlados	a) coletar e armazenar os metadados. b) verificar as palavras-chave que são idênticas com os termos dos vocabulários controlados. c) verificar as palavras-chave que remetem à igualdade dos termos do vocabulário controlado. d) vincular as palavras-chave e termos dos vocabulários controlados.
Tratar manualmente as palavras-chave não compatibilizadas automaticamente	a) verificar as palavras-chave que tenham termos correspondentes nos vocabulários controlados. b) inserir a palavra-chave como termo no vocabulário controlado da revista.

Fonte: Elaborado pelo autor (2020) fundamentado em Santos e Cervantes (2015b).

Santos, Cervantes, Londero e Goncalez (2016) tiveram como objetivo propor a implantação da VCPC *Tools* para o controle de vocabulário em periódicos científicos eletrônicos de áreas especializadas, nesse caso, o periódico Discursos Fotográficos, disponível no Portal de Periódicos Científicos da Universidade Estadual de Londrina, visando aprimorar a representação de conteúdos por meio da atribuição de palavras-chave. Os procedimentos metodológicos foram coleta de metadados e análises iniciais. Aplicaram-se quatro requisitos funcionais (coleta, análise, tratamento intelectual e resultados) mapeados conforme o ambiente do periódico científico. Considerou-se, neste reconhecimento, como as palavras-chave estavam atribuídas e, desse modo, como poderiam ser encontradas e quais as decorrências para a representação dos conteúdos temáticos.

Dentre os principais resultados, o trabalho traz um índice de termos que apresenta também a correlação deste com os respectivos artigos científicos eletrônicos. Rodrigues, Pereira, Londero e Cervantes (2017) trabalham a análise das palavras-chave do periódico Discursos Fotográficos coletadas a partir de Santos, Cervantes, Londero e Goncalez (2016). Contou-se com a revisão de especialista para consolidar as palavras-chave relevantes para descritores. Foram organizados os descritores em 12 classes e 12 subclasses, e o estudo dos autores focou na classe Fotografia. Algumas dificuldades foram encontradas devido à falta de instrumentos terminológicos documentários da área de Comunicação; neste sentido, foi importante e necessária a participação do especialista da área, executando os processos de revisão das palavras-chave por meio da análise das características terminológicas, para poderem ser consideradas descritores.

Santos, Cervantes e Londero (2018) apresentam um novo modelo de controle de vocabulário controlado para periódicos científicos eletrônicos, a partir da aplicação da VCPC *Tools* como aporte para desenvolvimento de um vocabulário controlado para a área da

Comunicação. A revista utilizada como estudo experimental foi a Discursos Fotográficos, vinculada ao Programa de Comunicação da Universidade Estadual de Londrina (UEL). A falta de instrumentos da área de domínio é uma das dificuldades para suprir o perfil de publicações de uma revista. O processo contém as seguintes etapas: identificar o endereço do repositório digital do periódico; desenvolver o projeto físico do banco de dados; elaborar a estrutura conceitual do domínio e subdomínio do periódico em estudo; levantar os instrumentos terminológicos do domínio e subdomínio; analisar as palavras-chave com os termos dos instrumentos terminológicos; verificar as palavras-chave que são idênticas entre si; verificar as palavras-chave que tenham termos correspondentes nos instrumentos terminológicos; classificar, tematicamente, as palavras-chave conforme a estrutura conceitual do periódico; inserir a palavra-chave como candidata a um termo no vocabulário controlado da revista.

Todo o processo foi realizado, inicialmente, de maneira semiautomática com a VCPC *Tools*. Logo após o término da execução, preparou-se o índice de termos para importação no TemaTres, o qual realizou com eficácia o processo de importação para implementar um modelo de interoperabilidade entre os sistemas. A VCPC *Tools*, que foi devidamente testada, recebeu em 2017 a licença-patente emitida pelo Instituto Nacional da Propriedade Industrial (INPI).

O vocabulário controlado entra em uso nos periódicos científicos eletrônicos, como instrumento de recuperação da informação, em conformidade com os princípios de Lancaster (1987), que são: representação consistente, por meio do controle dos quase sinônimos e sinônimos, e distinções de homógrafos, para facilitar a busca exaustiva por meio de agrupamento de termos com seus significados paradigmático e sintagmaticamente relacionados. O tesauro é uma variedade de vocabulário controlado que é aplicado normalmente para padronizar cabeçalhos de assuntos ou descritores, em qualquer circunstância que seja requerida uma padronização terminológica, adota-se o tesauro (LANCASTER, 1987).

Para a finalidade de obter uma maior eficiência nos processos de desenvolvimento de um vocabulário controlado, torna-se necessária a adoção de um *software* para a construção e manutenção. O mesmo vale para os metadados na recuperação em catálogos online com padrões descritivos e temáticos, os quais são utilizados na construção de registros bibliográficos e de autoridades. Um dos principais requisitos, em um *software* de gestão de tesauro, são as funcionalidades de construção e a possibilidade de manter, em constante evolução, a linguagem. Outro aspecto importante é a interoperabilidade do vocabulário controlado proporcionada pelo *software* em ambientes multilíngues (FUJITA *et al.*, 2017).

A finalidade do tesauro é servir de guia para o indexador e para o pesquisador na atribuição de um termo para representar, de maneira eficiente e eficaz, um conceito. O tesauro,

quando integrado em um sistema de busca, permite que os termos possam ser explorados por meio de cinco formas: expansão de busca; sugestão de termos de pesquisas alternativas; suporte para agrupamento ou filtros para uma busca; identificação de digitação; e suporte na indexação automática (ISO 25964-1).

Precisa-se pensar nos procedimentos que sistematizam todas as atividades fins para a construção e gestão deste vocabulário controlado. Nesta perspectiva, de integração da gestão do tesauro em sistemas computacionais, inserem-se os sistemas utilizados para este propósito. Fujita et al. (2017) apresentam uma análise da literatura que aponta para os sistemas MultiTes e o TemaTres e abordam aspectos relevantes desses sistemas, como os principais requisitos. A compilação de termos para produzir um tesauro é trabalho muito oneroso e é possível realizar sem um suporte de *software*, mas o uso de *software* de gestão de tesauros aumenta a eficiência nos processos de construção e evita erros que comprometam o instrumento (ISO 25.964-1).

O TemaTres é um *software* livre aplicado na gestão de representações linguísticas formais do conhecimento, desenvolvido por Diego Ferreyra em 2004 para ambiente *Web*. Dentre os instrumentos gerenciados estão: vocabulários controlados, taxonomias, tesauros, glossários e lista de cabeçalho de assunto (GONZALES AGUILAR; RAMÍREZ POSADA; FERREYRA, 2012; TEMATRES, 2018). No TemaTres, estão presentes os requisitos básicos da gestão de representações linguísticas formais do conhecimento como: inserção de termos, construção de relacionamentos como USE, UP, TG, TE, TR, cadastro de diversos tipos de notas (escopo, aplicação, bibliográfica), possibilidade de interoperabilidade por meio do *Simple Knowledge Organization System* (SKOS).

O TemaTres é descrito também no próximo capítulo, que traz a fundamentação do planejamento e estruturação de vocabulários controlados, tanto o processo de maneira intelectual quanto os processos aplicados para construção de vocabulários em periódico científico eletrônico. Nessa construção indica-se *a priori* a atuação dos padrões de Hearst (1992, 1998) para auxílio na construção da sistemática - hierarquia dos termos. Considera-se que as pesquisas com a temática vocabulários controlados em periódicos científicos eletrônicos ainda é, relativamente à relevância do tema e à quantidade de estudos já produzidos sobre vocabulários controlados, de modo geral, escassa. A necessidade de recuperar a informação com rapidez e precisão é, cada vez mais, fundamental como condição para a produção de novos conhecimentos. Não houvessem outros, isso seria motivo suficiente para o desenvolvimento de estudos atuais sobre vocabulários controlados.

3 ESTRUTURA SISTEMÁTICA DE VOCABULÁRIOS CONTROLADOS COM SUBSÍDIOS DA LINGUÍSTICA E DA APLICAÇÃO DE PADRÕES LÉXICO-SINTÁTICOS

Os aspectos teóricos e metodológicos da construção, manutenção e avaliação de SOC abrangem um conjunto de processos que fazem parte do núcleo de estudos da CI. Naturalmente, os elementos que integram o SOC são os conceitos - que são oriundos da teoria, mesmo com sua complexidade e dificuldade de identificação. O sentido expressado pelo conceito é fundamentado nas teorias, em outras palavras, conceitos são a matéria prima para constituição de uma estrutura conceitual (HJØRLAND, 2015).

As abordagens deste capítulo versarão sobre os aspectos linguísticos no que se refere ao SOC e ao subsídio da análise de *corpus* para desenvolvimento de SOC, recomendações dos processos de construção de SOC, com abrangência dos vocabulários controlados, sedimentados pelo tesauro que atualmente está formalizado pela norma ISO 25.964:1. Relacionam-se essas recomendações com os aspectos de padrões léxico-sintáticos como recurso na identificação de relacionamentos hierárquico.

3.1 ASPECTOS LINGUÍSTICOS NO VOCABULÁRIO CONTROLADO

As comunidades se estruturam, dentre outros aspectos, de uma organização social denominada língua, que por sua vez se desenvolve em um espaço de tempo, e atua lado a lado com a linguagem, definida como um mecanismo de representar conceitos por meio de uma língua ou outros sistemas de representação considerados mais amplos em relação à pura noção de língua. Neste contexto, a língua remete ao idioma como um sistema de convenção de uma comunidade de usuários que sofre constantes modificações; a linguagem são sistemas de representação de conceitos por meio de sinais, sons, imagens, conjuntos de caracteres, entre outros mecanismos de representação, como linguagem de programação, linguagem documentária, linguagem de indexação (ALVARES, 2012).

O vocabulário controlado pode ser entendido como uma interlocução entre os produtores da informação, os organizadores e os utilizadores da informação por meio de uma linguagem unificada de maneira artificial, tendo como base a linguagem natural. Em suma, é uma compatibilização das linguagens dos envolvidos no processo informacional intermediada pelos pontos de acesso - descritores (AGUIAR; TÁLAMO, 2012).

Neste sentido, os fenômenos linguísticos, que são característicos de cada cultura, tendem a afetar essa interlocução e interromper o ciclo comunicativo que ocorre desde o produtor, organizador e utilizador, influenciado na recuperação da informação. As características básicas da língua natural são sinonímia, polissemia e a homografia. A sinonímia é a possibilidade de um conceito ser representado por diversos termos. Seguindo o caminho inverso da sinonímia, a homografia e a polissemia têm capacidades semelhantes. A polissemia dá ao termo mais de uma acepção com um núcleo comum, sendo o significado diferente, porém com uma relação entre as diversas acepções. A homografia é a capacidade de palavras diferentes, que possuem a mesma escrita, poder nomear diversos conceitos distintos e sem relações. Ambas as variações linguísticas têm relação com o comportamento da língua, em campos semânticos próximos, regionalização, cultura, empréstimos de termos entre línguas, entre outras variações que influenciam os fenômenos linguísticos (ALVARES, 2012).

O vocabulário controlado é, portanto, a constituição de uma linguagem por meio de um conjunto de termos sintetizados de maneira não ambígua e não redundante. Recorre-se à área de conhecimento da linguística de *corpus* como metodologia teórico-prática de análise, subsidia as extrações de relações léxico-sintáticas (discutidas no item 3.3) como contribuição no desenvolvimento de SOC, considerando-se a complexidade da língua em seus aspectos culturais, regionais entre as demais variantes. Oliveira (2009, p.49) corrobora com Alvares (2012) no sentido de caracterizar a linguagem como um fenômeno social e o fato de que sua análise se dá por meio da comunicação, buscando significado no discurso. Integram na linguística as áreas de conhecimento que configuram uma interdisciplinaridade entre Linguística de *Corpus*, Linguística Sistêmico-Funcional (LSF), Linguística Aplicada (LA), Linguística Computacional (LC) entre outras. Para efeitos deste estudo, enfatizou-se a área de conhecimento da Linguística de *Corpus*.

A Linguística de *Corpus* tem como ocupação a coleta e a exploração de dados linguísticos textuais, denominados *corpora*, seguindo criteriosamente os procedimentos de coleta. Por conseguinte, analisa e pesquisa questões relacionadas a uma língua e as variantes linguísticas, em busca de evidências empíricas. Desde muito tempo, já existiam os *corpora*, que no sentido da palavra é *corpus* ou corpo. “Na Grécia Antiga, Alexandre, o Grande, definiu o Corpus Helenístico. Na Antiguidade e na Idade Média, produziam-se *corpora* de citações da Bíblia.” (SARDINHA, 2004, p.3).

A área vem se desenvolvendo desde os primeiros *corpora*, sendo o primeiro o *Brow Corpus*, na década de 60 do século passado, com 1 milhão de palavras em inglês americano. Um outro *corpus* é o *Lancaster-Oslo/Bergen Corpus* (LOB), em inglês britânico, e utilizado na

década de 1970. A grande expansão foi a partir de 1980, diante de aspectos favoráveis, como sócio-históricos, acadêmicos, tecnológicos e pragmáticos. Por outro lado, pesquisadores como Geoffrey Leech, Jan Svartvik, John Sinclair, Randolph Quirk e Douglas Biber foram os linguistas responsáveis pelo desenvolvimento e disseminação acadêmica, bem como as diversas possibilidades que os *corpora* podem evidenciar, desde estudos gramaticais e novas tendências da língua e uso (OLIVEIRA, 2009, p. 67). Para Carvalho (2007, p.15), "[...] a visão da linguagem como sistema probabilístico e a abordagem empirista, destacando a análise de dados provenientes da observação da linguagem, reunidos sob a forma de um *corpus* [...]" é a base da Linguística de *Corpus*. Somente nos anos 1990 esta área do conhecimento inicia, pela mão de pesquisadores interessados, o desenvolvimento de análise de *corpus* linguísticos em língua portuguesa, com isso, algumas iniciativas organizam os *corpora* do português (OLIVEIRA, 2009, p. 57).

Devido a essa expansão da área, a integração com a tecnologia pôde acelerar os fatores de desenvolvimento, com aportes de máquinas computacionais robustas com grande capacidade de armazenamento e processamento. Em contrapartida, para utilizar esses recursos de *hardware*, precisa-se de sistemas sofisticados para a manipulação e análise dos *corpora*, o que depende do desenvolvimento, de forma colaborativa, dessas propostas de soluções por uma equipe de pesquisadores interdisciplinar. A realidade da interdisciplinaridade ainda é apenas uma proposta e pode ser um aspecto de retardo da evolução da área (OLIVEIRA, 2009, p. 67). Para Carvalho (2007), é admirável a expressividade de pesquisadores que ainda trabalham em *corpora* não legíveis por máquina, uma vez que, diante de um contexto tecnológico, com o uso de um computador pessoal doméstico, é possível realizar a análise de milhares de palavras de maneira confiável e rápida. Em atividades repetitivas como a contagem de palavras, listagem de termos frequentes, a margem de erro é bastante reduzida.

O autor supracitado reconhece que:

> O papel do computador na ciência moderna já é bem conhecido. Mas no estudo da linguagem, os computadores têm oferecido novas perspectivas. Eles nos permitem ver fenômenos antes despercebidos devido às limitações da nossa cognição. O impacto do uso de *corpora* computadorizados na linguística tem sido associado ao uso do telescópio na astronomia ou ao microscópio na biologia. Mas o estudo da linguagem não teve o seu telescópio ou microscópio; o computador é a ferramenta analítica que torna possível, pela primeira vez, um verdadeiro estudo empírico da linguagem. Ou seja, a partir dos estudos linguísticos com base em *corpora* computadorizados, o uso e a frequência de palavras e expressões em uma determinada língua são comprovados através de sua ocorrência, tanto individualmente como em conjunto com outras palavras (coocorrência). (CARVALHO, 2007, p. 14)

Em meio aos argumentos em discussão das variáveis dessa área do conhecimento, surge a complexidade de definir a Linguística de *Corpus* como sendo uma metodologia ou disciplina que implica a definição dos termos metodologia e disciplina. A metodologia é entendida como somente instrumental, e a disciplina como orientação teórica, logo a Linguística de *Corpus* não é somente uma metodologia, e sim metodologia e disciplina, pois é possível aplicá-la como ferramenta e manter as orientações teóricas da disciplina. Por exemplo, teríamos a sintaxe baseada em *corpus* e a tradicional, para as quais os distanciamentos seriam "[...] o instrumental; os dados, orientação, pressupostos teóricos, as implicações dos resultados e tudo o mais permaneceriam". Se entender a metodologia "[...] como um modo típico de aplicar um conjunto de pressupostos de caráter teórico [...], assim se poderá visualizar a Linguística de *Corpus* como uma metodologia" (SARDINHA, 2004, p.36).

O autor elenca seis pontos importantes para garantir um *corpora* de qualidade:

> A origem: os dados devem ser autênticos.
> O propósito: o corpus deve ter a finalidade de ser um objeto de estudo linguístico.
> A composição: o conteúdo do *corpus* deve ser criteriosamente escolhido.
> A formatação: os dados do *corpus* devem ser legíveis por computador.
> A representatividade: o *corpus* deve ser representativo de uma língua ou variedade.
> A extensão: o *corpus* deve ser vasto para ser representativo.
> (SARDINHA, 2004, p.18-19)

Por outro lado, apresenta quatro pré-requisitos para a formação de um *corpus* computadorizado, que são:

> 1) O *corpus* deve ser composto de textos autênticos, em linguagem natural. Assim, os textos não podem ter sido produzidos com o propósito de serem alvo de pesquisa linguística, e não podem ter sido criados em linguagem artificial, tal como linguagem de programação de computadores ou notação matemática.
> 2) Autenticidade dos textos subentende textos escritos por falantes nativos. Tanto assim que, quando esse não é o caso, deve-se qualificá-lo como *corpora* de aprendizes (*learner corpora*).
> 3) O conteúdo do *corpus* deve ser escolhido criteriosamente. Os princípios da escolha dos textos devem seguir, acima de tudo, as condições de naturalidade e autenticidade. Mas devem também obedecer a um conjunto de regras estabelecidas por seus criadores de modo que o *corpus* coletado corresponda às características desejadas. Por exemplo, se é um *corpus* de português brasileiro escrito que represente a língua portuguesa, tal qual é escrita no Brasil, em sua totalidade, a coleta deve ser guiada por um conjunto de critérios que garanta, entre outras coisas, que o maior número possível de tipos textuais existentes no português brasileiro esteja representado, que haja uma quantidade aceitável de cada tipo de texto e que a seleção dos textos seja aleatória, a fim de não contaminar a coleta com variáveis indesejáveis.
> 4) Representatividade. Tradicionalmente, tende se a ver um *corpus* como um conjunto representativo de uma variedade linguística ou mesmo de um idioma.

> Mas a questão não pode ser enfocada no vácuo. Cabe perguntar: representativo do quê e para quem? (SARDINHA, 2004, p.19-20).

A anotação de um *corpus* é o enriquecimento com informações linguísticas adicionadas de maneira manual por humanos ou automática por meio de computadores para um fim teórico ou prático. Além do valor de um *corpus* para as análises das variantes linguísticas, se ainda tiver sido anotado, aumenta significativamente o quão valioso se torna este *corpus*. Em decorrência desta anotação, é possível realizar buscas mais precisas e processamentos refinados (PEDRO; VALE, 2018, p.23). Para o conceito de representatividade do *corpus*, além das características e pontos importantes, Sardinha (2004) complementa que não são expressamente ditos os critérios para obtê-las. Quando designa que um *corpus* é representativo, o entendimento está na extensão dele, a partir de uma quantidade de palavras e de textos. Considera-se a discussão em torno da Linguística de *Corpus* de extrema importância para subsídios de análise das variantes linguísticas, a fim de consolidar as estruturas de linguagens artificiais no campo da organização e representação do conhecimento, processos que passam por meio de recomendações de melhores práticas disseminadas na norma ISO 25.964:1.

3.2 Recomendações de processos de construção vocabulários controlados: ISO 25.964:1

Conforme a norma ISO 25.964:1, a construção de um tesauro é trabalho de esforço intenso. Algumas considerações iniciais sobre o tesauro devem ser delineadas como: será usado para quê e por quem; terá limitações do software que será utilizado; e qual o nível do conhecimento dos usuários. Os estágios iniciais da compilação compreendem no geral: quando e como começar; coleta de termos e conceitos; e análise de termos. No estágio de análise de termos coletados, eles devem ser organizados em ordem sistemática para dar suporte na tomada de decisão de inseri-los no tesauro. Esta organização por meio de planilhas compreendendo os assuntos ou facetas facilita a aproximação dos termos com a finalidade de verificar as variantes e sinônimos. Embora rudimentar, a classificação realizada não precisa determinar a estrutura do tesauro, neste estágio (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011).

A sequência de trabalho na construção do tesauro é voltada sempre para o agrupamento dos conceitos hierarquicamente. A recomendação para o esboço da estruturação é *top-down*, devido a abordagem *bottom-up* poder gerar anomalia na hierarquização dos conceitos em níveis superiores. Um esboço de estrutura de alto nível deve ser preparado antes que o trabalho

prossiga aos níveis mais específicos da hierarquia. Torna-se mais eficiente o trabalho com termos agrupados hierarquicamente, com a inserção de suas equivalências (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011).

A relação hierárquica pode ser estabelecida entre dois conceitos no caso de um deles estar completamente incluído no outro. Fundamentam-se em graus ou superordenação e subordinação, em que o conceito superordenado representa uma classe ou conceitos subordinados aos seus membros. As etiquetas de identificação são BT (*broader term*) e NT (*narrower term*), em português as etiquetas são nomeadas como TG e TE, respectivamente. São três os tipos de relação hierárquica: genérico; todo-parte; instância. Além das etiquetas BT e NT para nominar as relações hierárquicas, há as etiquetas BTG/NTG - identificar relações genéricas; BTP/NTP - identificar relações todo-parte; e BTI - identificar relações de instâncias (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2011).

Nesse contexto de superordenação e subordinação que caracteriza as relações hierárquicas, relaciona-se com esta organização a sistemática, que desde a Grécia antiga já se apresentava como forma de denominar os aspectos de um determinado conjunto; em outras palavras, um caráter classificatório. Portanto, trata-se de princípios e métodos para classificar. Nos aspectos linguísticos em torno da sistemática, o termo é enquadrado em duas classes de palavras: adjetivo e substantivo. Percebe-se o crescimento da sistemática com a informática, portanto em buscas pela internet, ela aparece no sentido de adjetivo integrado a um substantivo de ordenamento ou classificação. Em casos de aplicação como substantivo, a grande maioria apresenta-se nas áreas Biológicas. Ao unir a palavra informática, o sentido dos resultados trata, hipoteticamente, dos aspectos de uma taxonomia informática. Portanto a sistemática adjetivada vinculada a um substantivo caracteriza-se por realizar um ordenamento a partir de métodos, lógico e coerente, científico e aplicável (CURRÁS, 2010).

Na classificação, os conceitos de superordenação e subordinação, que formam a sistemática, é possível relacionar com a hiperonímia e hiponímia, que estabelecem posições de amplitude e especificidade em uma hierarquia, que Moreira, Santos e Vitorini (2017, p.166) descrevem etimologicamente os termos:

> Uma análise etimológica breve dos termos deixa ver claramente seus significados. O prefixo “hiper” diz respeito a algo que está em posição superior (ou em excesso), como ocorre, por exemplo, em “hipertensão” (def.: “Med. Pressão excessiva exercida pelo sangue nas paredes dos vasos sanguíneos”); o prefixo “hipo”, por sua vez, relaciona-se a algo que está em posição inferior (ou em escassez), como ocorre, por exemplo, em “hipotensão” (def.: “Med. Pressão do sangue nas paredes dos vasos sanguíneos inferior à normal; pressão baixa”).

> Assim, hiperonímia é a palavra que transmite o sentido do todo e hiponímia remete à ideia de parte, tipo ou item do todo. Um hiperônimo é um termo que está superordenado em relação a um hipônimo; este, por sua vez, está subordinado em relação ao hipônimo. Trata-se, como se vê, de uma relação assimétrica. O termo "árvores frutíferas" é hiperônimo de "macieiras", "abacateiros" e "mangueiras". Tomando-se a relação em sentido oposto, "macieiras", "abacateiros" e "mangueiras" são hipônimos de "árvores frutíferas". Trata-se também de um tipo de relação estrutural, ao mesmo tempo em que "árvores frutíferas" é hiperônimo de "macieiras" pode ser também hipônimo de "árvores".

Na caracterização da hiperonímia e hiponímia, incluem-se ferramentas lexicais-sintáticas que, por meio da identificação de padrões lexicais, auxiliam a tomada de decisão na consolidação de relações hierárquica. A partir de recursos tecnológicos aplicados no mecanismo de automatização de processos de localização de um conceito superordenado e subordinado, há a finalidade de auxílio para o profissional da informação. São incluídos os padrões de Hearst, mecanismo para extração de padrões léxico-sintáticos e consequentemente de relações hierárquicas.

3.3 PADRÕES DE HEARST: PRINCIPAIS ABORDAGENS

A professora doutora Marti A. Hearst é docente na *School of Information* e *EECS Departament (Eletrical Engineering and Computer Sciences)* na *UC Berkeley (University of California, Berkeley)* e a primeira pesquisadora em interfaces de usuário para mecanismos de buscas, visualização da informação, processamento de linguagem natural. Dentre seus estudos, estão o de 1992, que descreve métodos que identificam um conjunto de padrões léxico-sintáticos que aparecem de maneira desigual e por meio de delimitações de gêneros textuais, tendo como resultado a análise de um tesauro construído de maneira manual e aplicações na recuperação da informação. Conforme consulta ao perfil da referida pesquisadora no Google Acadêmico, o estudo de 1992 tem 3.729 citações (consulta em 13/05/2019). O estudo de 1998 está associado ao método de identificação de relações léxico-semânticas a partir de padrões léxico-sintáticos, com 496 citações no Google Acadêmico.

Dentre os autores que utilizam os estudos de Hearst (1992, 1998) estão: Morin (1999) - citado por Brewster e Wilks (2004), Brewster, Ciravegna e Wilks (2002) - citados por Brewster e Wilks (2004), Brewster e Wilks (2004), Gillam, Tariq e Ahmad (2005), Baségio (2007), Freitas e Quental (2007), Freitas (2007), Khosravi e Vazifedoost (2007), Torres (2012), Taba (2013), Taba e Caseli (2014), Machado (2015), Machado e Lima (2015), Moreira, Santos e Vitorini (2017), Cimiano, Mädche, Staab e Völker (2019), Nédellec, Nazarenko e Bossy

(2019). Vale destacar que dentre os autores mencionados, da CI são somente Moreira, Santos e Vitorini (2017), os demais estão concentrados, em sua maioria, na área da Ciência da Computação e Linguística.

Os padrões de Hearst são aplicados a grande base de dados disponibilizados na *Web*, *corpora*, tendo como processo o *Lexico-Syntactic Pattern Extraction* (LSPE). A proposta de Hearst é um mecanismo de auxílio para os lexicógrafos e engenheiros de sistemas de conhecimento. Cimiano, Mädche, Staab, and Völker (2019) e N´Edellec, Nazarenko e Bossy (2018) abordam o conceito de aprendizagem de ontologia e os padrões de Hearst (1992, 1998) são incluídos na discussão como processo para automatizar a construção de relações hierárquicas. Brewster e Wilks (2004) também utilizam os padrões de Hearst (1992) para identificar relações hierárquicas no processo de desenvolvimento de ontologias a partir de três processos: associar termos, construir hierarquias e rotular as relações. Deste modo, utilizam os padrões léxico-sintáticos como mecanismo automático ou semiautomático para a construção de relações hierárquicas em diversos sistemas de organização do conhecimento. Apresentam-se, no Quadro 4, os objetos e/ou SOC de estudo dos autores com os padrões de Hearst.

Quadro 4 - Aplicações padrões léxico-sintáticos

Ano	Autores	SOC/Objeto	Língua Aplicada
1992	Marti A. Hearst	Comparação dos resultados com Tesauro WordNet.	Inglês
1998	Marti A. Hearst	Coleta automática de Relações da WordNet.	Inglês
2002	Christopher Brewster Fabio Ciravegna Yorick Wilks	Metodologia centrada no usuário para construção de ontologias utilizando sistema Adaptativa.	Inglês
2004	Christopher Brewster Yorick Wilks	*Discussão teórica - corpora* - análise do Processamento de Linguagem Natural para construção de ontologias e taxonomias.	-
2005	Lee Gillam Mariam Tariq Khurshid Ahmad	Metodologia para extração hierarquias conceituais de coleções de textos de um domínio para construção de ontologias.	Inglês
2007	Túlio Lima Baségio	Extração de conceitos e relações taxonômicas, para subsidiar a construção de ontologias - língua portuguesa.	Português BR
2007	Maria Cláudia de Freitas	Subsídios para a elaboração automática, a partir de *corpus*, de ontologias específicas quanto ao domínio (língua portuguesa).	Português BR
2007	Maria Cláudia de Freitas Violeta Quental	Subsídios linguísticos para a construção automática de taxonomias a partir de um *corpus* (língua portuguesa).	Português BR

Continua

Continuação

2007	Fariborz Khosravi Alireza Vazifedoost	Reestruturação de tesauro ASFA em persa com métodos de aprendizagem ontológica para obtenção de uma ontologia.	Persa
2012	Carlos Eduardo Atencio Torres	Proposta de método para reduzir o processo de construção de ontologias por especialista por meio de aprendizado de ontologia supervisionado.	Português BR
2013	Leonardo Sameshima Taba	Extração automática de relações semânticas binárias em língua portuguesa.	Português BR
2014	Leonardo Sameshima Taba Helena de Medeiros Caseli	Extração automática de relações semânticas binárias em língua portuguesa.	Português BR
2015	Pablo Neves Machado	Extração de relações hiponímicas a partir de corpora em língua portuguesa.	Português BR
2015	Pablo Neves Machado Vera Lúcia Strube de Lima	Padrões para extração de relações hiponímicas com base em um *corpus* de língua portuguesa.	Português BR
2017	Walter Moreira José Carlos Francisco dos Santos Érica Fernanda Vitorini	Leitura Documentária em artigos científicos - Processos auxiliares por meio da identificação de relações hierárquicas.	Português BR
2019	Claire Nédellec Adeline Nazarenko Robert Bossy	*Discussão teórica* - Extração de informação como conhecimento baseado em Processamento de Linguagem Natural – ontologias.	-
2019	Philip Cimiano Alexander Mädche Steffen Staab Johanna Völker	*Discussão teórica* - Aprendizagem ontológica (Ontology Learning).	-

Fonte: Dados da pesquisa.

Na observação dos padrões de Hearst (1992, p.541) e dos exemplos é possível compreender melhor as relações conceituais de hiperonímia e hiponímia, apresentadas no Quadro 5.

Quadro 5 - Padrões léxico-sintáticos e exemplos

Nº	Padrão	Exemplo
1	NP_0 such as NP_1 {, NP_2…, (or | and) NP_i } for all NP_i, $i \geq 1$, hyponym(NPi , NP0)	Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use. hyponym(“Gelidium”,“red algae”).
2	such NP as {NP, }*{(or|and)} NP	... works by such authors as Herrick, Goldsmith, and Shakespeare. hyponym(“author”, “Herrick”), hyponym(“author”, “Goldsmith”), hyponym(“author”, “Shakespeare”)
3	NP {, NP}* {,} or other NP	Bruises, ..., broken bones or other injuries ... hyponym(“bruise”, “injury”), hyponym(“broken bone”, “injury”)
4	NP {, NP}* {,} and other NP	... temples, treasuries, and other important civic buildings. hyponym(“temple”, “civic building”), hyponym(“treasury”, “civic building”)
5	NP {,} including {NP,}*{or|and} NP	All common-law countries, including Canada and England ... hyponym(“Canada”, “common-law country”), hyponym(“England”, “common-law country”)
6	NP {,} especially {NP,}*{or|and} NP	... most European countries, especially France, England, and Spain. hyponym(“France”, “European country”), hyponym(“England”, “European country”) hyponym(“Spain”, “European country”)

Fonte: Elaborado pelo autor (2020) fundamentado em Hearst (1992, 1998).

Hearst (1992), ao comparar um hipônimo como resultado com a hierarquia da *WordNet*, aponta três tipos de resultados possíveis, assim denominados: Verificar, Crítica e Aumentar. O resultado “verificar” é etiquetado se os termos estiverem na *WordNet* e se a relação hiponímia estiver na hierarquia, então o tesauro será verificado. O resultado “crítica” configura-se se ambos, estiverem na *WordNet* e se a relação hiponímia não estiver na hierarquia, então o tesauro é criticado, isto é, um novo conjunto de conexões hipônimo é sugerido. O resultado “aumentar” é se um ou ambos os termos não forem encontrados no tesauro, então essas frases nominais e sua relação são sugeridas como entradas.

No Quadro 6 são apresentados os padrões originais e a versão adaptada/traduzida por Moreira, Santos e Vitorini (2017). As transições e adaptações dos demais autores constam no quadro completo nos apêndices.

Quadro 6 - Padrões léxico-sintáticos adaptados/traduzidos português

Padrão Hearst (1992, 1998)	**Moreira, Santos e Vitorini (2017)**
NP_0 such as NP_1 {, NP_2…, (or | and) NP_i } for all NP_i, $i \geq 1$, hyponym(NPi, NP0)	SN (tais como | como) SN { , SN ... , } (e | ou) SN
such NP as {NP, }*{(or|and)} NP	tal(is) SN como {(SN,)*(ou|e)} SN
NP {, NP}* {,} or other NP	SN {, SN}* {,} ou outro(s) SN
NP {, NP}* {,} and other NP	SN {, SN}* {,} e outro(s) SN
NP {,} including {NP,}*{or|and} NP	SN {,} incluindo {SN,}*{ou|e} SN
NP {,} especially {NP,}*{or|and} NP	SN {,} especialmente {SN,}*{ou|e} SN (6b) SN {,} principalmente {SN,}*{ou|e} SN (6c) SN {,} particularmente {SN,}*{ou|e} SN (6d) SN {,} em especial { SN,}*{ou|e} SN (6e) SN {,} em particular { SN,}*{ou|e} SN (6f) SN {,} de maneira especial { SN,}*{ou|e} SN (6g) SN {,} sobretudo { SN,}*{ou|e} SN

Fonte: Elaborado pelo autor (2020) adaptado de Moreira, Santos e Vitorini (2017).

Os padrões de Hearst são considerados, para efeitos deste estudo, um mecanismo de potencialidade de localização de relações hierárquicas, por meio dos padrões léxico-sintáticos para auxiliar a compatibilização sistemática. Este potencial pode ser reconhecido a partir das características que Hearst (1998, p. 4) relaciona, sendo: a) “eles ocorrem frequentemente em muitos gêneros textuais”; b) “eles (quase) sempre indicam a relação de interesse”; c) “eles podem ser reconhecidos com pouco ou nenhum conhecimento pré-codificado”. A aplicação com traduções e adaptações em outros idiomas é relativamente possível, como já aplicado por Freitas (2007), Freitas e Quental (2007), Baségio (2007), Torres (2012), Taba (2013), Taba e Caseli (2014), Machado (2015), Machado e Lima (2015) e Moreira, Santos e Vitorini (2017) em língua portuguesa do Brasil.

A construção deste capítulo teve a finalidade de observar os aspectos da construção de vocabulários controlados, seguidos de conceitos da linguística e sua área de conhecimento de Linguística de *Corpus*, que realizam a coleta e a exploração de conjuntos de dados linguísticos, a fim de evidenciar, de maneira empírica, os fenômenos de uma língua, passando pelas recomendações da norma ISO 25.964:1. O contexto de identificação de relacionamentos léxico-

sintáticos por meio dos padrões de Hearst (1992, 1998) é tópico que, em síntese, vai ser descrito por meio de modelos de dados no instrumento de controle de vocabulário para constituir a interoperabilidade entre outros vocabulários, conforme a segunda parte da norma ISO 25.964-2, assunto em evidência no próximo capítulo.

4 MODELO E LINGUAGEM DE INTEROPERABILIDADE ENTRE VOCABULÁRIOS CONTROLADOS

As diversas designações para o conceito geral de interoperabilidade conduzem para um mesmo sentido: estabelecer *a priori* uma "ponte" entre sistemas, instrumentos, linguagens, entre outros. A palavra "interoperabilidade" sugere um termo atual e integrado com a tecnologia, porém é possível observar os métodos aplicados há vários anos em unidades de informação. Registra-se a abordagem dos diversos conceitos encontrados na literatura que cercam a interoperabilidade: conversibilidade, compatibilidade, mapeamento, intercâmbio, integração, alinhamento, união, combinação. Perpassam camadas como a sintática, semântica, estrutural e sistêmica. Logo, para obter um mapeamento e a compatibilização entre os instrumentos, torna-se necessário um modelo de dados que representa esse instrumento que será interoperado com outros instrumentos. Portanto é fato o estabelecimento de uma representação a partir do modelo de dados, a fim de parear um diálogo, neste contexto insere-se o SKOS.

4.1 CONCEITOS FUNDAMENTAIS DE INTEROPERABILIDADE

A interoperabilidade, *grosso modo*, é caracterizada pela possibilidade de concretizar a comunicação de sistemas computacionais por meio de uma linguagem comum entre eles. Para esse acontecimento, tornam-se necessários protocolos com regras lógicas elaboradas para tratar, tanto a codificação quanto a decodificação. Por codificação compreendem-se os processos de empacotar a informação para o transporte, sendo a decodificação o seu desempacotamento. Campos (2007, p.23) define a interoperabilidade como "[...] a habilidade para transferir e utilizar informações entre sistemas com eficiência e uniformidade, exigindo padronização e flexibilidade em certo nível." (CAMPOS, 2007, p. 23). Cunha e Cavalcanti (2008, p.213) conceituam-na como a "[...] capacidade que possuem os computadores de fabricantes distintos de trabalharem juntos usando um conjunto comum de protocolos para a comunicação e troca de informações dentro de uma rede".

A interoperabilidade não é um conceito atual, existe desde meados do século XX. Frente à explosão informacional, internet, rede de computadores, as bibliotecas já realizavam troca de informações, serviços cooperados, ordenamento universal, mas com a consolidação da biblioteca digital, a interoperabilidade passa a ter uma ênfase mais explícita. A partir da

arquitetura dos sistemas distribuídos e de editoração científica, cresce a necessidade de um modelo de negócio para a disseminação dos produtos em repositórios digitais. Aplica-se, nessas situações, a interoperabilidade, o que justifica a importância dada a ela nesse campo (SAYÃO; MARCONDES, 2008).

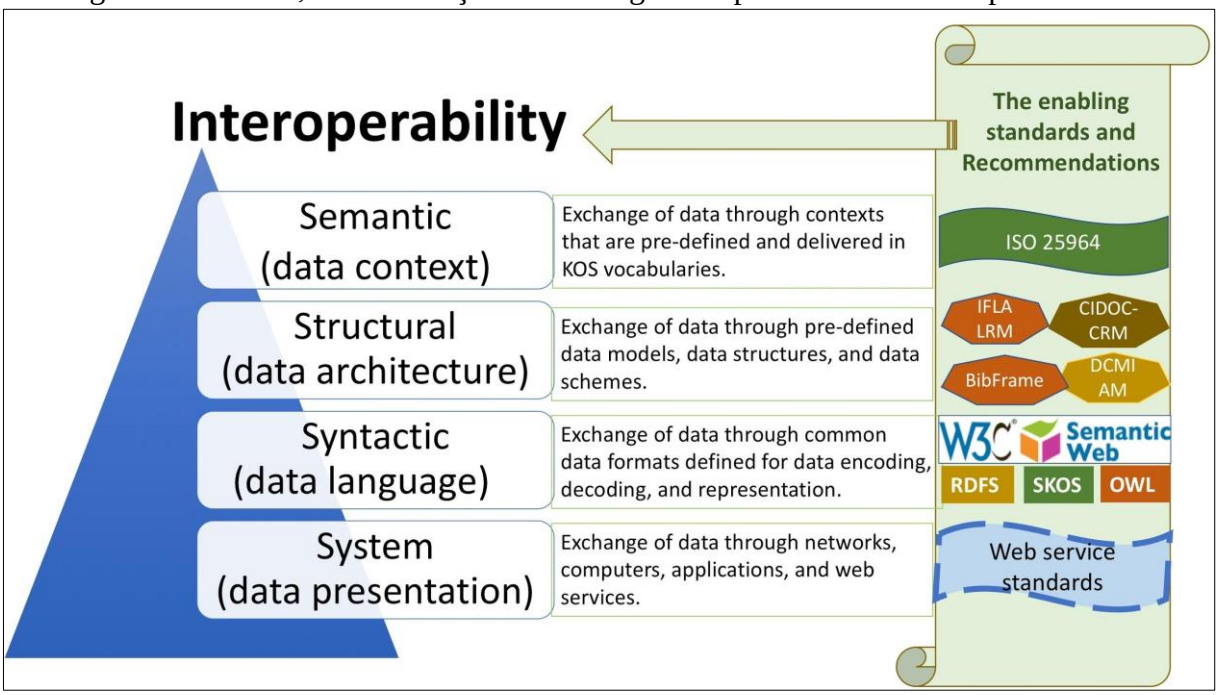
Neste contexto interoperável, o *Big data* tem influenciado no SOC, como Ibekwe-San e Bowker (2017) abordam em seu estudo, ao afirmar que implicações conceituais, epistemológicas e metodológica do *Big data* poderiam influenciar no processo de construção e manipulação de SOC. Diante de uma crescente desmaterialização física para era digital, o SOC pode ter seu ciclo de vida diminuído ou alterado. Além desse fenômeno algorítmico, em que a humanidade está o tempo todo conectada e gerando dados, vale inferir a relação com o paradigma colaborativo da *Web 2.0*, fator importante e analisado pelos profissionais da organização do conhecimento.

A interoperabilidade é definida como "Capacidade intrínseca de dois ou mais sistemas da organização do conhecimento ou sistemas de informação, de compartilhar, trocar e procurar dados ou informações" (BARITÉ, 2015, p.89). A necessidade de compartilhar, recuperar e catalogar informações e recursos de maneira relevante e significativa ressalta a importância dos níveis de interoperabilidade, no conceito de reutilização, entre os sistemas que auxiliam no processo de organização do conhecimento. Os níveis de interoperabilidade técnica, semântica e organizacional são suportados na qualidade dos sistemas de subsidiar e permitir o intercâmbio de informação (MOREIRO GONZÁLEZ; SÁNCHEZ CUADRADO; MORATO LARA, 2012).

A norma ANSI/NISO Z39.19-2005 (R2010) apresenta, como principal finalidade da interoperabilidade, o desenvolvimento de relacionamento, com detalhamento suficiente, entre vocabulários controlados construídos de modo individualizado. Essa possibilidade de vínculo entre os vocabulários controlados é voltada para a utilização em base ou banco de dados, para facilitar a recuperação da informação com sucesso. Desse modo, os resultados recuperados da pesquisa de recurso por meio de diferentes vocabulários controlados, devem ser semelhantes em ambos os vocabulários.

Zeng (2019) traz em seu estudo quatro aspectos de heterogeneidade da interoperabilidade que também acabam configurando-se, simultaneamente, como quatro problemas. Fundamentado nos autores Ouksel e Sheth (1999); Adebesin *et al*. (2013); Obrst (2003); Ontolog (2018) apresenta esses quatro aspectos como recomendações de interoperabilidade, definindo-os como camada, sendo elas: sistema, sintática, estrutural e semântica. Na Figura 4 apresenta-se a ilustração de Zeng (2019).

Figura 4 - Padrões, recomendações e abordagens de problemas da interoperabilidade



Fonte: Zeng (2019, p. 124).

Na camada de sistemas, alguns dos problemas são a incompatibilidade de *hardware* e sistemas operacionais, no quesito da troca técnica de dados entre aplicativos, sistemas de informação, sistemas *Web*, entre outros que compõem a rede de informação local e *Web*. Na camada sintática, as indagações relacionam-se com os padrões de codificação e decodificação de dados, dificultando a interoperabilidade. Por outro lado, a *World Wide Web Consortium* (W3C) traz recomendações para utilização de um padrão de linguagem. A camada estrutural tem suas implicações relacionadas à disposição arquitetural da informação, estrutura de dados, base de dados, modelos de dados, esquemas de dados, os quais afetam a interoperabilidade. As comunidades LAM (*library, archive and museum*) é uma das principais organizações que instituem recomendações como propostas de solução. Já a camada semântica é a mais complexa, pois que não se limita somente às variações linguísticas, espacial, temporal, terminológicas, relacionamentos, entre outros aspectos. Neste quesito, a norma ISO 25.964:2 configura melhores práticas para minimizar problemas de interoperabilidade (ZENG, 2019).

A integração do SOC na era *big data* é uma oportunidade de expressar a importância da base de conhecimento proporcionada pelos instrumentos em uma aplicação avançada. Essas aplicações de SOC podem ser elencadas: a) base de conhecimento para questões-respostas e computação cognitiva; b) dados ligados; c) sistemas de informação empresariais; d) interoperabilidade dos dados de sistemas de informação; e) interoperabilidade e reutilização de dados; f) base de conhecimento para extração de informações; g) diagramas de vínculos de nós

(mapas conceituais, mapas causais, diagramas de processos); h) organização do conhecimento para entender e aprender; e i) transferência de conhecimento entre domínios. (SOERGEL, 2015).

Em termos de vocabulário controlado conceitua-se a interoperabilidade como qualidade de dois ou mais SOC ou componente de informação intercambiar informações. Essa troca de informações entre o SOC é possível a partir do mapeamento e da compatibilidade entre formatos padronizados e protocolos computacionais compatíveis para realizar a interoperabilidade. Portanto estas características principais do SOC têm como finalidade oferecer a interoperabilidade (ISO 25.964:2).

Nesse sentido, Zeng e Chan (2004) apresentam oito procedimentos de interoperabilidade[6]: 1) Derivação/Modelagem; 2) Tradução/Adaptação; 3) Satélite e Nós ou Vinculação de União; 4) Mapeamento direto; 5) Mapeamento de coocorrência; 6) Comutação/Linguagem intermediária; 7) Vinculação por meio de uma lista de união temporária; e 8) Vinculação por meio de um protocolo de servidor de tesauro. Referidos precedimentos são descritos no Quadro 7. Svenonius (2010) caracteriza como métodos e fatores de interoperabilidade, e uso de múltiplos vocabulários os seguintes aspectos[7]: 1) Mapeamento direto; 2) Comutação de vocabulário; 3) Fatores para a interoperabilidade bem sucedida de vocabulários; 4) Mapeamento semântico; 5) Interoperabilidade idiomática (questões de terminologia multilíngue e idiomas dominantes); 6) Vocabulário satélite e de extensão.

Quadro 7 - Procedimentos de interoperabilidade

Processos metodológicos	**Descritivo**
1) Derivation/Modeling	Vocabulários mais específicos desenvolvidos ou derivados a partir de outro vocabulário.
2) Translation/Adaptation	Vocabulários desenvolvidos a partir da tradução ou adaptação de um vocabulário em outro idioma que compartilha da mesma estrutura e normas.
3) Satellite and Leaf Node Linking	Satélite: Vocabulário constituído a partir de parte da macroestrutura de outro vocabulário. Nós ou *linking*: nós hierarquicamente organizados de um vocabulário fonte é a base para originar ou vincular com outros vocabulários específicos.
4) Direct Mapping	Estabelecimento de relações de equivalência entre os termos de diferentes vocabulários.
5) Co-occurrence Mapping	Grupos de termos relacionados a partir da coocorrência.

Continua

[6] *1) Derivation/Modeling; 2) Translation/Adaptation; 3) Satellite and Leaf Node Linking; 4) Direct Mapping; 5) Co-occurrence Mapping; 6) Switching; 7) Linking Through a Temporary Union List; e 8) Linking Through a Thesaurus Server Protocol*.

[7] *1) Direct mapping; 2) Switching Vocabulary; 3) Factors for Successful Interoperability of Vocabularies; 4) Semantic Mapping; 5) Interoperability across Languages (Issues of Multilingual Terminology e Dominant Languages); 6) Satellite and Extension Vocabularies.*

Continuação

6) Switching	Vocabulário intermediário constituído a partir de outros vocabulários, com a finalidade de ligar um vocabulário ao outro.
7) Linking Through a Temporary Union List	*Software* de estabelecimento de correspondência entre termos de diferentes vocabulários (não necessariamente equivalem ao mesmo conceito) extraídos a partir de retorno de consultas de usuários. A apresentação se dá por meio de lista de união temporal e instancia um novo vocabulário.
8) Linking Through a Thesaurus Server Protocol	Sistematização de consultas a vários webservices de tesauros e demonstração dos resultados, sem a necessidade de um novo tesauro.

Fonte: Zeng e Chan (2004).

Boccato e Torquetti (2012, p.85) sintetizam que as dificuldades encontradas para obter a interoperabilidade na construção de linguagens são os tratamentos dos conceitos e termos em diferentes linguagens, e apresentam com base em Maniez (1997) e Lancaster (2002) os seguintes tratamentos:

> a) especificidade: enquanto uma linguagem de indexação pode ser bem detalhada em sua precisão terminológica, outra pode conter apenas termos gerais para descrever o mesmo objeto;
> b) exaustividade: enquanto uma linguagem omite alguns conceitos de um campo científico, a outra abrange todas as suas manifestações;
> c) termos compostos: decisão de pré-coordenar ou pós-coordenar os termos compostos, cabendo a cada instituição fazer sua escolha, de acordo com a política de indexação do sistema;
> d) sinônimos: decisão do termo preferido a ser eleito para representar tematicamente ("traduzir") um conteúdo documentário. Essa ação decisória é influenciada pela região em que é elaborada e utilizada a linguagem de indexação, enfatizada pelo princípio da garantia cultural;
> e) relações e sistemas de coordenação: as relações hierárquicas, as significações dos termos, bem como os seus sistemas de coordenação (pós-coordenado e pré-coordenado) podem ser diferentes entre linguagens, revelando estruturas distintas de construção. Cada instituição elabora a sua linguagem de acordo com o seu objetivo, focada no ambiente em que ela está inserida, considerando o acervo documentário, os profissionais da informação, os usuários, o sistema de recuperação da informação, entre outros, podendo, assim, variar os níveis hierárquicos e o repertório terminológico, representativos das diversas instituições, elementos ratificados pelas garantias literária, de uso e organizacional.

Barité (2015, p.89-90), por seu turno, resume os sete fatores que interferem na interoperabilidade, a partir da norma ANSI/NISO Z39.19-2005 (R2010):

> a) Similaridade entre conteúdos temáticos nos diferentes domínios;
> b) Diferentes vocabulários controlados usados para indexar conteúdo de domínios semelhantes;
> c) Especificidade ou granularidade gradual dos vocabulários controlados usados para indexar conteúdos diferentes de domínios ou bancos de dados;
> d) Diferenças no tratamento de sinônimos e quase sinônimos;
> e) Metodologias de busca exigidas pelas bases de dados utilizadas;
> f) Garantias (literárias, de usuário e organizacionais) utilizadas no desenvolvimento do vocabulário;
> g) Objetivos pretendidos pelos responsáveis pelos bancos de dados e sistemas.

Neste contexto, vale ressaltar a unificação de SOC x dados, SOC x SOC. Unificação define-se por utilização de mesmos tipos de entidades e tipos de relacionamentos para definir um conceito de forma clara, ou destacando suas diferenças. Por outro lado, a utilização de mesmo termo para os conceitos ou o mapeamento de termos que definem determinado conceito significa a unificação. A unificação envolve cruzar fronteiras, onde os limites da interoperabilidade e troca sejam permeáveis. No que tange a essa unificação, surge a flexibilização que dá a conotação de adaptação e não apenas tradução, em um ambiente de variações linguísticas e culturais. A integração do SOC pode ser descrita em três momentos: horizontal em mesmo domínio, horizontal em domínios diferentes e vertical (SOERGEL, 2015).

A finalidade da interoperabilidade entre vocabulários, para a recuperação da informação, é permitir que a expressão de busca formulada seja convertida ou complementada por meio de uma expressão correspondente em outros vocabulários controlados (ISO 25.964:2). Andrade (2015) aborda em sua pesquisa que o termo interoperabilidade é encontrado na literatura especializada como conversibilidade, compatibilidade e mapeamento. Hinz e Pallazo (2008), Daltio (2007), Felicíssimo (2004) tratam processos de interoperabilidade em ontologias na área da computação como intercâmbio, integração, mapeamento, alinhamento, união, combinação. Sabe-se que em suas pesquisas o foco está na ontologia e na área da computação. No Quadro 8 apresentam-se, para efeitos desta tese, os conceitos de interoperabilidade, mapeamento e compatibilidade a partir de Barité (2015) e suas variações do conceito na literatura.

Quadro 8 - Conceitos interoperabilidade

Termo (BARITÉ, 2015)	**Etapas do processo**
Interoperabilidade EN: *Interoperability* ES: *Interoperabitidad*	Estrutural, Sintática, Semântica e Sistêmica (ZENG, 2019) Técnica, Semântica e Organizacional (MOREIRO GONZÁLEZ; SÁNCHEZ CUADRADO; MORATO LARA, 2012) Técnica, semântica, organizacional, política e humana, intercomunitária, legal e internacional (ANDRADE; CERVANTES, 2012)
Mapeamento EN: *Mapping* ES: *Mapeo*	Conversibilidade (LANCASTER; SMITH, 1983) Estudo dos tipos de incompatibilidades (NEVILLE, 1970)
Compatibilidade EN: *Compatibility* ES: *Compatibilidad*	Reconciliação de tesauro (NEVILLE, 1970) Técnica (LANCASTER; SMITH, 1983, LANCASTER, 1995; 2002) Estrutural (LANCASTER, SMITH, 1983, LANCASTER 1995;2002, HUDON, 2004) Lexical (HUDON, 2004) Compatibilidade semântica (NEVILLE,1970) Compatibilidade verbal (DAHLBERG, 1981) Compatibilidade de conceitos (DAHLBERG, 1983; HUDON, 2004) Compatibilidade de assuntos (HUDON, 2004; RIESTHUIS, 1996)

Fonte: Dados da pesquisa.

O conceito mapeado em um ou vários vocabulários controlados pode aumentar a quantidade de termos correspondentes ou que representem um mesmo conceito. No trabalho realizado pela interoperabilidade, é possível o estabelecimento de mapeamentos interconceitos e particularmente equivalência (ISO 25.964:2).

No contexto da recuperação da informação, há dois estágios principais nos quais os mapeamentos entre vocabulários controlados podem ser usados: (a) como parte do processo de indexação ou (b) no momento da pesquisa. Porém estas não são as únicas situações em que os mapeamentos podem ser usados e as recomendações da ISO 25.964:2 não devem ser interpretadas como excludentes para outras aplicações (ISO 25-964:2).

Nos mapeamentos usados no processo de indexação, o índice de termos dos metadados de documentos indexados com o Vocabulário A são convertidos para os termos correspondentes do Vocabulário B. Este processo pode ser realizado rotineiramente ou na atualização de uma grande coleção, ou em casos de pequenas atualizações sempre que os vocabulários são atualizados e/ou novos documentos são adicionados à coleção. A coleção de documentos pode então ser explorada ou pesquisado usando o Vocabulário B por um período indefinido, sem necessidade de mapeamento adicional (ISO 25.964:2).

O termo mapeamento é um referencial aos processos de estabelecimento de relações entre os conceitos de um vocabulário controlado com outro externo. A abordagem do mapeamento é disseminada para almejar a interoperabilidade semântica entre vocabulários de SOC. O mapeamento pode se referir a um produto do processamento do mapeamento, como declarações de relações entre termos, notações ou conceitos de um vocabulário relacionado a um outro. O vocabulário de SOC é caracterizado pelas instâncias de SOC como forma de diferenciá-lo dos vocabulários de metadados. Na terminologia das comunidades de *Linked Open Data* (LOD), o vocabulário de SOC é nomeado como vocabulário de valor. Essa implicação do mapeamento ocorre por conta da independência na sua concepção (ZENG, 2019).

Em casos de aplicação dos mapeamentos aos termos de pesquisa, os metadados permanecem inalterados. Para obter a mesma capacidade (a de usar o vocabulário B para pesquisar uma coleção indexada com o vocabulário A), os vocabulários de origem e os de destino dos mapeamentos precisam ser estabelecidos na direção inversa. Isso permite que as consultas de pesquisa que compreendem termos do Vocabulário B sejam convertidas nos termos correspondentes do Vocabulário A. Assim os mapeamentos são incorporados ao processo de pesquisa e precisam ser aplicados toda vez que uma pesquisa é feita usando o Vocabulário B. (ISO 25.964:2).

Por mapeamento denomina-se uma técnica na OC que se destina a analisar um domínio e desenvolver uma representação gráfica de conceitos e relacionamentos. Esta representação pode ser expressa por meio de diagramas, mapas conceituais, mapas tópicos entre outros instrumentos (BARITÉ, 2015). O mapeamento, também nominado metaforicamente como *crosswalks*[8], é de extrema importância para a proposição da interoperabilidade em um contexto contemporâneo. Na tentativa de superar a heterogeneidade semântica, em ambientes com muitas variáveis terminológicas, essa travessia remete a uma busca integrada entre os diversos vocabulários controlados aplicados na indexação de assuntos (KEMPF; TITZE; ECKERT; ZAPILKO, 2014). O conceito de "*crosswalks*" também é definido na norma ISO 25.964:2 como tabela de mapeamento entre conceitos em dois ou mais vocabulários controlados e oferece a possibilidade de pesquisa em diferentes bancos de dados.

Por outro lado, as heterogeneidades, dentre elas o idioma, assunto/domínio e a estrutura do vocabulário, podem afetar a qualidade do mapeamento, sendo essa uma prática para obter a interoperabilidade. Torna-se importante a atenção voltada para os aspectos de heterogeneidades que mais afetam o mapeamento e a possível perda de informações. Os mapeamentos são caracterizados em três principais quesitos: equivalência, hierárquico e associativo. O mapeamento hierárquico se subdivide em geral e específico. A relação de qualidade no mapeamento está vinculada à prioridade no estabelecimento do mapeamento na ordem: equivalência (simples ou composto), hierárquico e associativo (JIA; ZHAO, 2015; ISO 2011; ANDRADE; LARA, 2016; KEMPF; TITZE; ECKERT; ZAPILKO, 2014).

O mapeamento mais necessário e recomendado é o de equivalência, o qual é estabelecido entre os sinônimos a partir da manifestação dos conceitos em dois ou mais SOC, sem existirem diferenças entre os conceitos, termos ou notações que representam (ANDRADE; LARA, 2016; KEMPF; TITZE; ECKERT; ZAPILKO, 2014). A integração do conceito de mapeamento com a compatibilidade relaciona-se por meio de processos sequenciais, ou seja, a compatibilidade acontece após o mapeamento. Os autores, porém, remetem intrinsecamente ao mapeamento nos processos de compatibilização, como o caso da matriz de reconciliação dos casos de Neville (1970), a matriz de compatibilização de Soergel (1974) e Dalhberg (1981).

A compatibilidade é uma qualidade de um sistema permitir que seus elementos sejam utilizados, de maneira complementar, por outro sistema por meio da interoperabilidade

[8] Conforme o *Cambridge Dictionary* nomina um local em uma estrada que o tráfego de veículos é interrompido para possibilitar a passagem de pedestres. Logo esta tradução não remete ao mapeamento, mas entende-se que é um mecanismo de ligação, como é definido na norma ISO 25.964:2. Por outro lado, literalmente, pode ser entendido como caminho cruzado, onde há uma intersecção é o ponto de conexão entre os objetos que necessariamente precisam da troca de informações para obter o êxito no seu percurso.

(DAHLBERG, 1983; BARITÉ, 2015; UNISIST, 1971). Nesta perspectiva, não seria necessário nenhum processo de conversão para a obtenção de compatibilidade, desconsiderando as diferenças de notação, estrutura e suporte físico (UNISIST, 1971).

A compatibilidade técnica determina o formato de registro e convenções na codificação, em base de dados. Diante dessa aplicação, duas ou mais base de dados poderiam ser, tecnicamente, compatibilizadas, tornando uma opção mesclada, por meio da qual poderia ser realizado diversas opções de buscas. Já em base de dados indexada por vocabulário controlado a complexidade de compatibilização é maior devido à possibilidade de um conceito ser representado de forma muito diferente. Neste caso, torna-se necessário o mapeamento ou conversibilidade entre os vocabulários controlados. Esse mapeamento determina que cada termo do vocabulário B equivalente no vocabulário A deve ser registrado a partir de seu vínculo. As equivalências podem ser desenvolvidas em formato legível por máquinas, o que poderá gerar um mapeamento de maneira automatizada. Em contribuição à compatibilização estrutural, as normas visa facilitar o mapeamento de um vocabulário com o outro, para o tesauro tem como referência as normas ISO. Os mapeamentos de um vocabulário A para B é diferente do inverso, portanto, caso tenha necessidade do mapeamento de B para A deve ser desenvolvido um mapeamento recíproco (LANCASTER; SMITH, 1983).

Considerando a utilização de vários vocabulários controlados em uma base de dados, para cada mapeamento entende-se como importante manter um mapeamento recíproco em ambiente no qual o usuário poderá realizar buscas por ambos os mapeamentos. O número de operações está relacionado com a quantidade de vocabulários na seguinte proporção: 2 vocabulários para 2 operações, 3 vocabulários para 6 operações e 4 vocabulários para 12 operações. Alternativamente a este processo de alto custo, pode ser desenvolvida uma linguagem neutra como forma de mapeamento entre os vocabulários controlados. Oportunamente, quanto maior a quantidade de instituições cooperadas, um modelo torna-se mais econômico comparado aos vocabulários bilaterais e aos mapeamentos (LANCASTER; SMITH, 1983).

Lancaster e Smith (1983), Lancaster (1995) e Lancaster (2002) fazem uma contextualização histórica dos pesquisadores que reproduziram estudos de compatibilização de tesauros, conforme listados no Quadro 9, acompanhado de uma síntese e o ano do estudo.

Quadro 9 - Síntese de Lancaster sobre os estudos de compatibilização

Ano	Autor	Síntese
Painter	1963	Grupo de pesquisadores que estudaram a possibilidade de converter a linguagem de indexação usada por uma agência do governo dos Estados Unidos para utilização em outra agência. Produziram tabelas de correspondências entre os termos, resultados que publicaram nos relatórios da *Datatrol Corporation*.
Jaster	1963	
Datatrol Corporation	1963	
Hammond e Rosenborg	1962 1964	
Hammond	1965	
Henshaw	1967	Analisa a conversibilidade dos esquemas de classificação bibliográfica em vocabulários.
Neville	1970	Apresenta alguns problemas de conversão de vocabulários e os níveis de correspondência.
Ranganathan e Neelameghan	1970	Estudo preliminar da compatibilidade da *Colon Classification* e da *Universal Decimal Classification*.
Smith	1974	Estudo sobre os problemas de conversão humana de vocabulário por meio de normas.
Protapova	1975	Estuda a utilização da compatibilidade do índice de assuntos usado pela *All-Union Book Chamber* (União Soviética), análise entre cinco bibliotecas.
Gopinath e Prasad	1976	Estudo sobre o grau de compatibilidade estrutural entre tesauros e esquemas de classificação.
Sobolevskya	1976	Estuda a utilização da compatibilidade do índice de assuntos usado pela *All-Union Book Chamber* (União Soviética), análise em uma biblioteca pública.
Höpker	1976	Trata do desenvolvimento de um tesauro médico no idioma Alemão compatível internacionalmente, com a *Clinical Key of Diagnosis* (KDS), a *International Classification of Diseases* (ICD), e a *Systematized Nomenclature of Pathology* (SNOP).
Antopol'skii e Rudykina	1977	Estuda a possibilidade de compatibilidade entre os tesauros existentes na União Soviética.
Dahlberg	1977	Analisa a conversibilidade dos esquemas de classificação bibliográfica em vocabulários.
Glushkov, Skorokhod'ko e Strognii	1978	Apresentam estudos sobre a distinção das compatibilidades: semântica e estrutural em linguagens de indexação.
Jachowicz	1979	Descreve a construção do *Polish Thesaurus of Hospital Science*, destacando a necessidade de torná-lo compatível com o *German Thesaurus Krankenhauswesen*, entre outros tesauros polonês.
Mishin	1979	Analisa a conversibilidade dos esquemas de classificação bibliográfica em vocabulários.
Sager	1981	Aborda o problema da conciliação de tesauros em Ciências Sociais.

Fonte: Elaborado pelo autor (2020) fundamentado em Lancaster e Smith (1983), Lancaster (1995) e Lancaster (2002).

No contexto da interoperabilidade de vocabulário controlado, vale referenciar dois métodos importantes para a compatibilização, são eles: o Método de Reconciliação de Tesauro, proposto por Neville (1970; 1972), e a Matriz de Compatibilização Conceitual de Dahlberg

(1981; 1983), que são citados pelos autores Campos (2005), Campos (2006), Campos *et. al* (2009), Boccato e Torquetti (2012) e Teixeira e Souza (2013).

Campos (2005, 2006), Campos *et. al* (2009), Boccato e Torquetti (2012), Teixeira e Souza (2013) argumentam que o princípio do método de Neville (1970, 1972) é a compatibilização pelo conceito e não pela expressão linguística que nomeia o conceito, fornecido por descritor. Estabelecem a relação de equivalência a partir de códigos numéricos atribuídos aos conceitos, por meio da reconciliação, e limitando a quantidade de descritores incompatíveis. O método de Dahlberg (1981, 1983) é abordado pelos autores Campos (2005), Campos (2006), Campos *et. al* (2009) e Teixeira e Souza (2013) como um mapeamento semântico entre as linguagens, o qual gera resultados da compatibilidade semântica e estrutural dos instrumentos.

Em outras palavras, Neville (1970) trata em seu estudo os tipos de incompatibilidades para realizar a conversão de termos de tesauros em outros sistemas. As incompatibilidades podem apresentar um número limitado e possível de gerenciar. A partir dessa gestão, é possível encontrar soluções para cada tipo de incompatibilidade e possibilitar desenvolvimento de um esquema de reconciliação ou de conversão para tesauros. Dessa forma, trata-se de uma solução geral para a reconciliação de tesauros e não um problema semântico de cada termo a ser tratado. Neste esquema de reconciliação, apresentam-se os casos que refletem os tipos de incompatibilidades, sendo 11:

- Caso 1: Correspondência exata entre os termos dos tesauros.
- Caso 2: Diferentes sinônimos são usados para um mesmo conceito.
- Caso 3: O tesauro de origem tem um termo para o conceito e no outro tesauro não existe.
- Caso 4: O termo no tesauro de origem existe no outro tesauro, mas com um termo mais geral.
- Caso 5: O tesauro de origem usa termos pré-coordenados e outro tesauro usa termos pós-coordenados.
- Caso 6a: O tesauro de origem distingue homônimos e o outro tesauro não.
- Caso 6b: O tesauro de origem não distingue homônimos e o outro tesauro faz distinção.
- Caso 7: um tesauro utiliza descritores separados para distinguir um termo usado em sentidos diferentes, enquanto o outro tesauro não o faz.

- Caso 8: o tesauro fonte utiliza descritores que não são suficientes para esclarecer o conceito a que se referem.
- Caso 9: o tesauro fonte contém descritores sinônimos.
- Caso 10: o tesauro fonte faz uso de termos com significado apenas para o uso no local de origem.
- Caso 11: um tesauro utiliza um sistema de codificação arbitrário para alguns conceitos.

No Quadro 10 apresenta-se um caso de compatibilização exemplo da matriz de correspondência de Neville (1970), no qual é atribuído um código-chave para o termo reconciliado a partir de um tesauro de origem.

Quadro 10 - Caso 1 modelo de correspondência simples

	Tesauro de Origem	Tesauro B	Tesauro C
Entrada original	AIRFIELDS	AIRFIELD	FLUGPLÄTZE
Entrada reconciliada	AIRFIELDS (0101)	AIRFIELD (0101)	FLUGPLÄTZE (0101)
Código-chave	0101= AIRFIELDS	0101= AIRFIELD	0101 = FLUGPLÄTZE

Fonte: Elaborado pelo autor (2020) adaptado de Neville (1970).

Posteriormente, Neville (1972) apresenta a aplicação do método de reconciliação proposto em 1970 em casos reais de tesauro. Foram escolhidos dois tesauros em dois idiomas diferentes, sendo um em inglês e outro em francês. Essa reconciliação, considerando a época do estudo, tinha como finalidade formar um instrumento de indexação de maneira cooperada para a área da ciência da construção. A reconciliação é definida como um método de atribuição de códigos numéricos para os termos dos tesauros que representam um mesmo conceito (Quadro 10). Isto é, em um contexto em que duas ou mais organizações tratam um mesmo conjunto literário, é possível utilizar tesauros diferentes para indexar seu acervo, sem a necessidade de reindexação para um intercâmbio de materiais.

Soergel (1974) apresenta um quadro de equivalências para mapeamento de categorias entre linguagens diferentes (Quadro 11). Esse quadro descreve três graus de equivalência: S1 – correspondência exata; S2 – Correspondência aproximada e S3 – Sem correspondência. O S2 é subdividido em três subníveis: S2.1 – Correspondência mais geral; S2.2 -Correspondência mais específica; e S2.3 – Correspondência relacionada.

Quadro 11 - Tabela modelo de mapeamento de conversibilidade

Expression of searching equivalente: to the A-descriptor corresponds		Degree of equivalence				
		(S1) Precise searching equivalente in B	(S2) Approximate searching equivalente in B			(S3) No searching equivalente in B
			(S2.1) Searching equivalente broader	(S2.2) Searching equivalente narrower	(S2.3) Searching equivalente related	
(a) One B-descriptors	(a1) Same term	A: Ship B: Ship	A: Radar (for traffic control) B: Radar		A: Furniture (meaning office) B: Furniture (home)	A: Tanks (combat vehicles_ B: Tanks (containers)
	(a2) Different terms	A: Ship B: Vessel	A: Military aircraft B: Airchaft ---------- A: Pesticides B: Pest control		A: Office Furniture B: Home furniture	A: Traffic simulation B: -
(b) A combination of B-descriptors	(b1) OR-combination of B-descriptors	A: Aircraft B: Aircraft OR Airplanes OR Helicopters OR ... ---------- A: Passenger transport B: Passenger cars OR Busses OR Passenger ships OR Passenger aircraft	A: Heavy weigth cargo transport B: Cargo cars OR Cargo Ships OR Cargo aircraft OR Cargo airplanes OR Cargo helicopters	A: Pesticides B: Herbicides OR Insecticides OR Rodenticides (Molluscacides omitted from B)		-
	(b2) AND-combination of B-descriptors	A: Rolling stock B: Vehicles AND Rail transport	A: Locomotives B: Vehicles AND Rail Transport		A: Astrology B: Stars AND Popular ideas	-
	(b3) Combination of B-descriptors using both AND and OR	A: Animal food products B: Food AND (Hunting OR Animal husbandry)	A: Meat as food B: Food AND (Hunting OR Animal husbandry)			-

Fonte: Soergel (1974, p. 506)[9].

[9] Optou-se por não traduzir o cabeçalho da tabela, a fim de evitar possíveis adaptações inadequada ao modelo.

Dahlberg (1981) estabelece quatro matrizes de compatibilização de linguagens de indexação, são elas: M1, M2, M3, M4. A M1 é uma matriz de organização e comparação alfabética dos termos, para obter a compatibilidade verbal. A M2 é a reorganização da M1 adicionando, em cada termo, o descritor ou classe. Na M2, podem haver diferenças entre os termos, sendo indicados com os seguintes sinais matemáticos: '≠' divergência de conceito; '<' conceito mais geral em relação a coluna 2 do Quadro 12; '>' conceito mais específico em relação a 2 do Quadro 12; 'c' combinações de conceitos. A M3 é uma reformulação alternativa à M2, baseada em um sistema mestre para instanciar as classes e descritores que se correlacionam com as linguagens de indexação, pode ser denominada como matriz de correlação. A M4 é um refinamento da M3, com a indicação de sistemas mais amplos comparados à linguagem de indexação mais específica.

No Quadro 12 apresenta-se exemplo da compatibilização conceitual estabelecida por Dahlberg (1981), na coluna 6 indica-se o número de coincidência conceitual.

Quadro 12 - Exemplo da matriz de compatibilização conceitual

Nº	Name	DDC	BBC	UNT	CC
65	Social Welfare	361 Social problems and social welfare	Q Social Welfare	R85/99 Social Welfare	3
65.1	Soc. welf. philos.	361.01 Philos. & theory of	QAE Philos. of soc. welf.	R86 Soc. welfare philosophy	3
65.2	Soc. welf. admin.	-	QAG Soc. administration, Soc. welf. administration	R90 Sociat welfare administration	2
65.3	Welfare policy	> 361.25 Action within establ. soc. framework (policy)	QAG P Policy (in Social Welfare)	R87 Welfare policy	3
65.4	Soc. welf. planning	> 361.25 Act. within establ. soc. framework (planning)	QAH Planning for welfare, social planning	R88 Social welfare planning	3
65.5	Soc. welf. econom.	-	QAT M/Z Management of soc. welf.	R89 Social welfare economics	2 2

Fonte: Dahlberg (1981, p.89).

Dahlberg (1983) complementa o estudo proposto da matriz de compatibilização de Dahlberg (1981) com ênfase na compatibilização conceitual em sistemas ordenados. Os sistemas ordenados nomeados são todo instrumento de organização, descrição e indexação, como tesauros, sistemas de classificação, lista de cabeçalhos entre outros. Instancia-se em três

níveis de compatibilidade conceitual: coincidência conceitual, conceitos semelhantes e correlação conceitual.

Lancaster e Smith (1983), Lancaster (1995) e Lancaster (2002) apresentam o algoritmo de conversão de tesauro dos pesquisadores Wall e Barnes de 1969. Esse algoritmo é reproduzido no Quadro 13 com base em Lancaster (1995).

Quadro 13 - Algoritmo de conversão de tesauro de Wall e Barnes

Processo de conversão	**Exemplo**
Correspondência exata	CALORIMETRÍA *en* CALORIMETRÍA
Variações ortográficas	ELECTRO OSMOSIS *en* ELECTROOSMOSIS HAEMOPHILUS *en* HEMOFILO
Forma escrita	REGISTRO ELÉCTRICO *en* REGISTRO ELECTRÓNICO
Inversões	PRESA, AVES DE *en* AVES DE PRESA
Referência cruzada	GRAN MAL *en* EPILEPSIA (donde un vocabulario contiene una referencia GRAN MAL *use* EPILEPSIA)
Hierarquia	SENO CAROTIDEO *en* ARTERIAS (dónde la jerarquía en el vocabulario origen indica que SENO CAROTIDEO está subordinado a ARTERIAS)

Fonte: Lancaster (1995, p.204).

A modelagem de SOC, de maneira a usar os mecanismos de interoperabilidade, torna indispensável um modelo de dados para a descrição de SOC com a finalidade de transcender os limites geográficos, quando se fala da disponibilização em ambiente *Web*, para estabelecimento da interoperabilidade.

4.2 MODELO DE DADOS E INTEROPERABILIDADE

No momento contemporâneo, muitos conceitos concorrem com a tecnologia. Aliás nada diferente para o termo interoperabilidade, que se vincula imediatamente com a existência de tecnologia da informação. A relação dos dois conceitos dependentes, utilizados por entidades como a W3C, instancia padrões para representação de SOC por meio do SKOS, ou seja, SKOS é um modelo de dado para representar SOC na perspectiva de intercâmbio de dados.

Os esquemas conceituais podem ser interligados quando disponíveis na *Web* por meio do SKOS, que por sua vez é apresentado como uma aplicação em *Resource Description Framework* (RDF). O SKOS é aplicado para mapear vocabulários controlados com a finalidade de representar conceitos, utilizando-se de rótulos ou etiquetas que o nomeiam, nesse contexto, é possível a troca de informações com outros esquemas conceituais e padrões de metadados (RAMALHO, 2015). O W3C define o RDF como um padrão de troca de dados na *Web*. Catarino, Cervantes e Andrade (2015) argumentam que "[...] o SKOS escrito em RDF, tornam os esquemas de conceito expressos neste modelo, passíveis de serem lidos por agentes inteligentes conforme as recomendações da W3C" (p. 110).

Nessa perspectiva de representação computacional de SOC por meio do SKOS no suporte RDF, Pastor-Sánchez e Martínez-Méndez (2009) definem o SKOS como um vocabulário em RDF para representar um SOC. Entre eles estão os tesauros, taxonomias, sistemas de classificação e lista de cabeçalhos de assunto. Essa descrição em SKOS/RDF surge a partir da necessidade de traduzir ou representar para leitura dos sistemas computacionais e possibilitar a interoperabilidade entre os diversos SOC. Pastor-Sánchez e Martínez-Méndez (2009) traduziram para o idioma espanhol o documento "SKOS Simple Knowledge Organization System" da W3C (ISAAC; SUMMERS, 2009). Para efeitos de exemplificação, na Figura 5 apresenta-se uma tripla em RDF. A codificação <skos:prefLabel> e <skos:altLabel> sinaliza o nome do termo ou etiqueta para o termo preferido e o segundo é um termo alternativo/sinônimo.

Figura 5 - Tripla em RDF

```
<A> rdf:type skos:Concept ;
  skos:prefLabel "love"@en ;
  skos:altLabel "adoration"@en ;
  skos:broader <B> ;
  skos:inScheme <S> .

<B> rdf:type skos:Concept ;
  skos:prefLabel "emotion"@en ;
  skos:altLabel "feeling"@en ;
  skos:topConceptOf <S> .

<S> rdf:type skos:ConceptScheme ;
  dct:title "My First Thesaurus" ;
  skos:hasTopConcept <B> .
```

Fonte: W3C (MILES; BECHHOFER, 2009a).

A partir dos estudos de Moreira (2012) que despertam a lacuna na literatura quanto a estudos e à disseminação de discussões relacionadas à consolidação teórica da proposta do

SKOS no período de 2006 a 2011, Santos e Moreira (2018) realizaram um levantamento em bases de dados científicas representativas da área da CI, tendo como finalidade levantar, categorizar e demonstrar as aplicações do SKOS relacionadas aos vocabulários controlados (Quadro 14). As bases utilizadas foram: *Information Science and Technology Abstracts* (ISTA), *Library Information Science Abstratcs* (LISA), *Library, Information Science & Technology Abstracts* (LISTA) e *Scopus*, totalizando 232 registros.

Quadro 14 - Análise dos artigos da categoria "aplicação – vocabulários controlados"

AUTOR	OBJETO	LOCAL AMBIÊNCIA	AGENTE	CAUSA E EFEITO
Binding & Tudhope (2016)	Vocabulários arqueológicos de diferentes países.	Getty Vocabulary LOD SPARQL	Projeto ARIADNE	Mapeamento de vocabulários
Caracciolo et al. (2012)	Tesauro de Agricultura	AGROVOC	Food and Agriculture Organisation (FAO)	Alinhamento com outro SOC, abordagem de direções atuais e futuras
Cohen & Franke (2015)	Vocabulário Militar	C2 ASCA DART	-	Interoperabilidade entre os vocabulários
Gray, Gray, Hall & Ounis (2010)	Vocabulário Controlado Astronomia	Vocabulary Explorer Vocabulary Explorer Web	Virtual Observatory	Recuperação de recursos por meio de vocabulários
Jia & Wei (2012)	Tesauro	Chinese Thesaurus (CT) Library of Congress Subject Headings (LCSH)	Scientific and Technical Information Institute of China – National Library of China	Mapeamento entre SOC
Nicholson & McCulloch (2006)	Descrição de assunto	Project HILT	Joint Information Systems Committee (JISC)	Mapeamento diferentes assuntos e esquemas de classificação
Papadakis & Kyprianos (2011)	Tesauro para sistema IR	Library of Congress Subject Headings (LCSH)	-	Mapeamento LCSH com tesauro para Integração em sistema IR
Wright, Harrison & Watkins (2015)	Tesauro dados de química ambiental	CEH Analytical Services Thesaurus (CAST)	Laboratory Information Management System (LIMS)	Interoperabilidade semântica entre os recursos etiquetados, mapeamento de demais SOC
Zapilko, Schaible, Mayr & Mathiak (2013)	Tesauro Ciências Sociais	Thesaurus for the Social Sciences (TheSoz)	Leibniz Institute for the Social Sciences (GESIS)	Mapeamento do tesauro em SKOS - LOD

Fonte: Santos e Moreira (2018).

No Quadro 14 apresenta-se a análise realizada nos trabalhos recuperados categorizados como aplicação em vocabulários controlados, e especificamente trata de mapeamento,

interoperabilidade e recuperação. O quadro completo com os demais trabalhos categorizados consta nos apêndices.

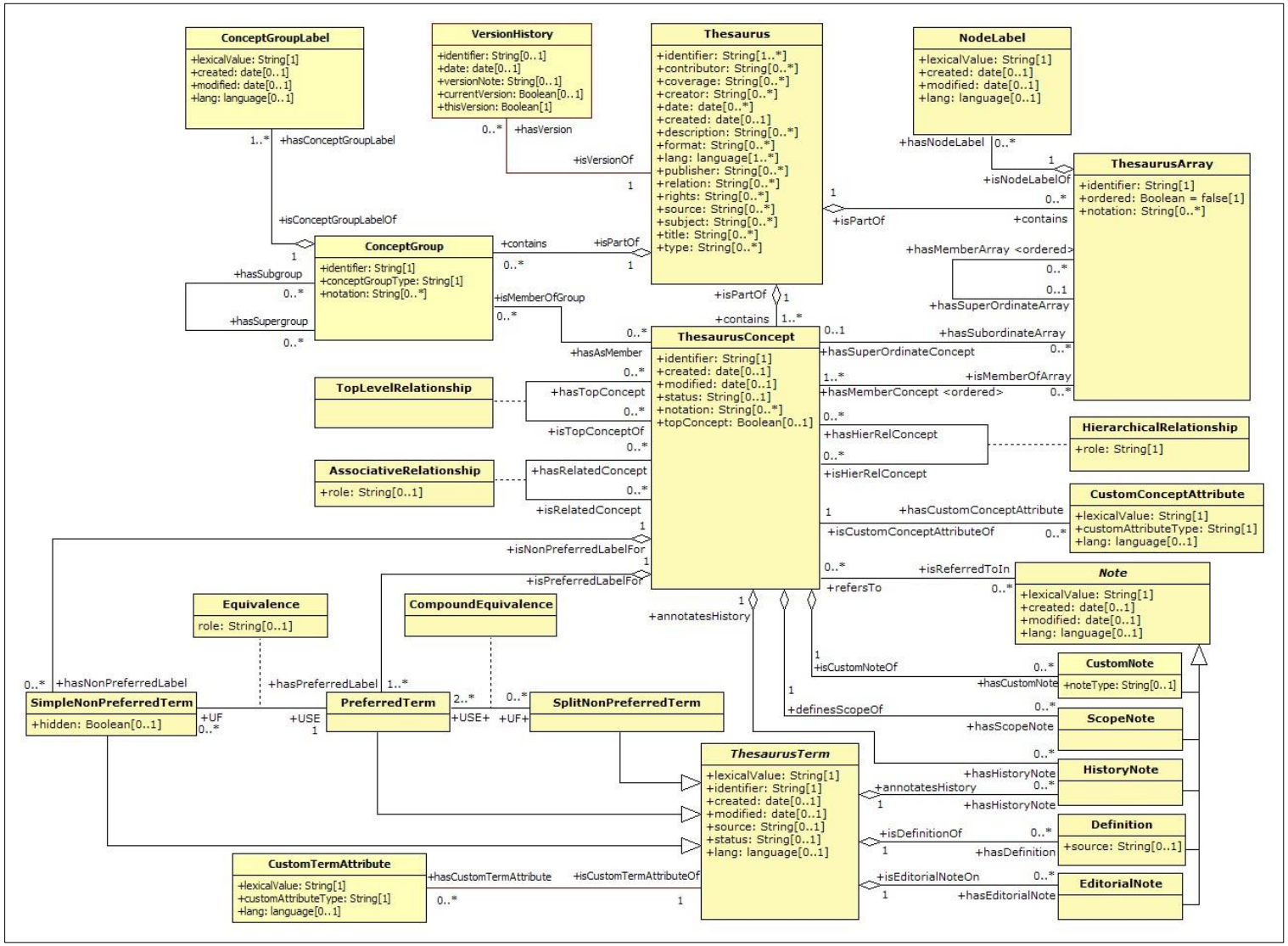
Santos e Moreira (2018) ressaltam a predominância dos trabalhos de SKOS relacionados com estudos de casos de interoperabilidade entre SOC a partir da sua publicação na *Web*. Observam também a integração de SKOS com o LOD no desenvolvimento de propostas de SOC interoperável e ainda com a possibilidade de estar aberto e ligado na *Web*. Uma diversidade de termos para o conceito de interoperabilidade é evidenciada, entre eles o termo integração, alinhamento e superação de barreiras linguísticas.

Ainda no contexto de modelo de dados, Coppens, Debevere e Mannens (2014, p.8) apresentam o diagrama de classe desenvolvido pela norma ISO 25.964:1 (p. 104), em que a referência do *data model* (Figura 6) deixa clara a finalidade de relacionar-se com outros vocabulários controlados, operacionalizando a interoperabilidade. Coppens, Debevere e Mannens (2014) esclarecem o termo-conceito em relação ao requisito ambiguidade. Para contemplar o requisito de ambiguidade, a estrutura dos vocabulários controlados é baseada em conceitos ao invés de termos. Isso significa que o modelo para descrever o vocabulário controlado mostra as relações entre os conceitos, e distingue-os dos termos que são utilizados para rotular esses conceitos. Desta forma, um vocabulário controlado pode lidar com a ambiguidade. Diferentes conceitos podem ser ligados aos mesmos termos. Por exemplo, o termo "Java" pode estar relacionado com três conceitos diferentes: ilha, café e linguagem de programação. Portanto os relacionamentos em um vocabulário controlado estão entre os conceitos, não os termos.

O modelo de dados é a explicitação de toda estrutura do vocabulário controlado modelado utilizando o diagrama de classe da *Unified Modeling Language* (UML) (Figura 6), deixa legível para a máquina o termo e o conceito, além de suas relações de associações, agregações e generalizações. Isso se deve à estrutura lógica do vocabulário controlado representado posteriormente em um modelo de dados como o SKOS, descrito em uma linguagem para ocorrer a interoperabilidade. Clarke e Zeng (2012) fazem uma discussão do termo e do conceito em um modelo de dados legível por computadores. O conceito é compreendido, intuitivamente, pelo humano, independentemente de palavras ou linguagem, mas para comunicá-lo é preciso representá-lo de alguma forma (palavra, figura, códigos). O termo é utilizado para representar o conceito normalmente em sistemas de busca e em tesauros. Apesar da confusão que permeia o conceito/termo, até o final do século XX não era reconhecida a importância do modelo de dados.

Neste contexto, esta falta de importância do modelo de dados decorre do fator humano controlar e gerenciar os instrumentos e intermediar o processo de representação, busca e recuperação. A partir da inclusão de máquinas para interpretar os dados é que a necessidade ganha destaque. Se não existe um modelo de dados, torna-se impossível a compreensão não só de termo e conceito, mas de qualquer estrutura do vocabulário controlado, consequentemente não existe a possibilidade de ocorrer a interoperabilidade. *A priori*, a contribuição do modelo de dados (Figura 6) é no sentido de definir padrões e proporcionar a interpretação de dados por meio de máquinas computacionais (CLARKE, ZENG, 2012).

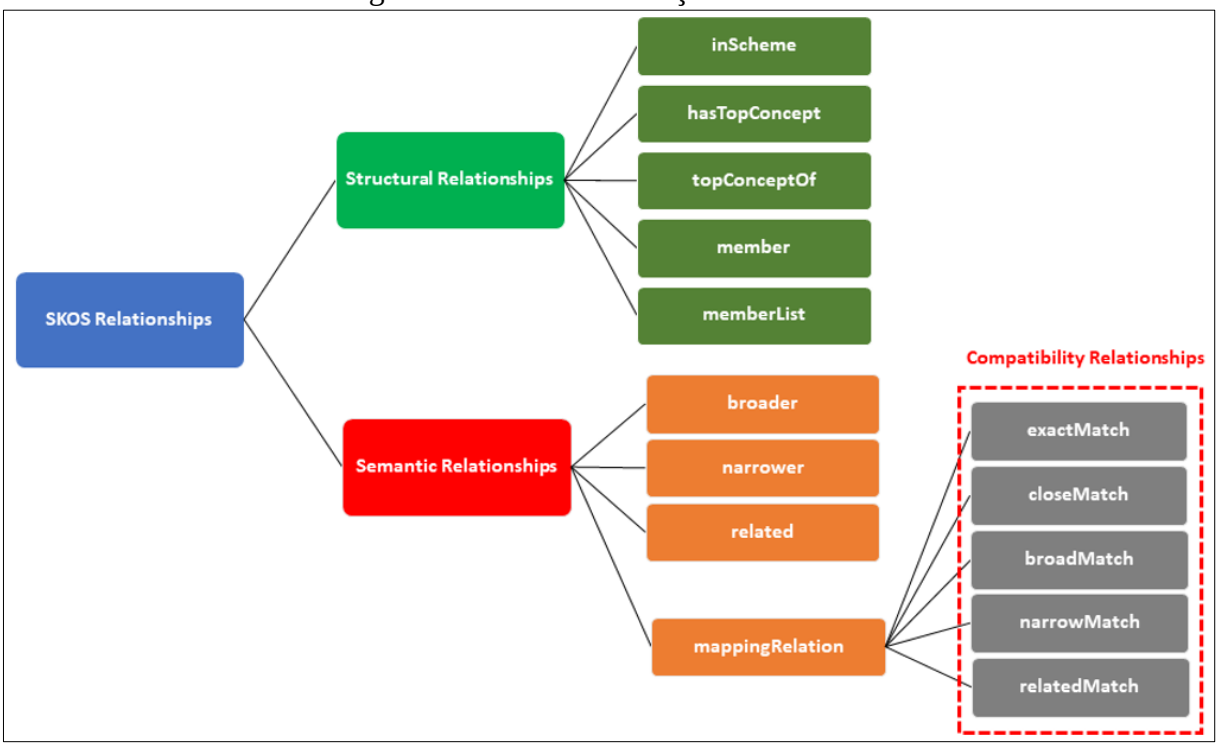
Figura 6 - Data Model - Modelo de dados



Fonte: ISO 25.964:1 (2011, p.104), Coppens, Debevere e Mannens (2014, p. 4).

O SKOS não se limita a somente representar os conceitos do SOC, mas integra a representação de coleções e esquemas de conceitos (RAMALHO; CERVANTES, 2019). Na Figura 7 são ilustradas as relações estruturais e semânticas, com um destaque para as relações de mapeamento.

Figura 7 - Síntese de Relações em SKOS



Fonte: Ramalho e Cervantes (2019, p. 6).

Na classificação de Ramalho e Cervantes (2019) as relações de mapeamento têm como função a representação da compatibilidade entre diferentes esquemas conceituais.

Em síntese, este capítulo apresentou os aportes teórico-práticos que permeiam a interoperabilidade, os conceitos que a definem como processo de compatibilidade entre o emissor e o receptor da informação, com destaque para o fato de que ambos necessitam de regras bem definidas, espelhando-se na teoria da comunicação. Mencionou-se também que existem outros termos que remetem a esse sentido, lembrando que a interoperabilidade sempre existiu, porém não era denominada com este termo. Um destaque foi para o fato de que é mais comum, no atual momento, esse termo relacionado com a tecnologia da informação, para efeitos deste estudo com o SOC. As reflexões perpassam pelo principal modelo de dados de representação de SOC o SKOS recomendado pela W3C e descrito em RDF, por onde se sistematiza a interoperabilidade, orientado pela norma ISO 25.964:2 e pelas boas práticas descritas sobre esse aspecto.

5 PROPOSTA TEÓRICO-METODOLÓGICA DE INTEROPERABILIDADE

A proposta teórico-metodológica fundamenta-se nos métodos e procedimentos de Santos (2015) conjugados às ações com base no objetivo geral desta tese. Os objetivos de Santos (2015) foram: desenvolver a ferramenta VCPC *Tools*; coletar e armazenar os vocabulários controlados; compatibilizar as palavras-chave com os termos dos vocabulários controlados; tratar manualmente as palavras-chave não compatibilizadas automaticamente. Portanto Santos (2015) propôs a compatibilização de palavras-chave atribuídas aos artigos científicos como suporte para a construção do vocabulário controlado do periódico.

A relação com o estudo realizado por Santos (2015) é demonstrada pelos aspectos de continuidade da pesquisa. De maneira sumária (para informações mais detalhadas consulte as subseções 5.1, 5.2, 5.3, 5.4, 5.5), a proposta desta tese é a apresentação de mecanismos complementares para o mapeamento e compatibilização de palavras-chave e o modelo de interoperabilidade entre os vocabulários controlados dos periódicos científicos. No Quadro 15, apresentam-se as ações geradas pelos objetivos específicos deste estudo.

Quadro 15 - Ações decorrentes dos objetivos específicos

Etapas	**Ações**
Reconhecer o ambiente de vocabulário controlado de periódicos científicos eletrônicos gerenciados pelo OJS. Etapa 1	1. Buscas em bases de dados e de teses e dissertações. 2. Buscas em currículos *lattes* – autores com a temática. 3. Visitas aos portais de periódicos com a finalidade de análise do VC. 4. Registro da disponibilização dos VC do periódico científico aos usuários.
Sistematizar o conteúdo dos artigos do periódico científico eletrônico em formato de busca em texto completo a partir da base de dados da VCPC *Tools*. Etapa 2	1. Coleção dos artigos definidos como *corpus* a partir das ações da Etapa 1. 2. Conversão em formato TXT. 3. Desenvolver projeto conceitual lógico do banco de dados. 4. Estruturar em linguagem SQL os arquivos TXT e inserir no banco de dados.
Identificar os padrões de Hearst a partir das palavras-chave não compatibilizadas. Etapa 3	1. Estruturação de consultas e localização de cada padrão de Hearst. 2. Análise de palavras anteriores e posteriores aos padrões. 3. Organização dos resultados e convalidação especialista da sistemática.

Continua

Continuação

Mapear, por meio dos padrões de Hearst (1992; 1998), a sistemática das palavras-chave com o vocabulário controlado e compatibilizar com os níveis de correspondência Etapa 4	1. Registro do relacionamento da palavra-chave e termo. 2. Inserção de estrutura sistemática do termo. 3. Verificação da interoperabilidade destes termos com o vocabulário de referência. 4. Instanciação de vínculos com vocabulário de referência.
Desenvolver o mapeamento reverso das palavras-chave compatibilizadas pelos processos de Santos (2015) a partir dos níveis de correspondência. Etapa 5	1. Modelagem e execução do mapeamento reverso de palavras-chave compatibilizadas por Santos (2015) 2. Desenvolvimento do modelo conceitual e lógico de interoperabilidade. 3. Organização dos termos para importação no vocabulário controlado integrativo dos periódicos. 4. Registro da interoperabilidade dos termos com o vocabulário de referência.

Fonte: Elaborado pelo autor (2020).

Na etapa 1, concentrou-se na proposta de identificar, a partir das bases de dados, BDTD e *Lattes*, os trabalhos relacionados com a implantação de vocabulários controlados em periódicos científicos eletrônicos para organizar uma lista de revistas que apresentassem esse tipo de organização e representação do conhecimento, com o foco naquelas que são gerenciadas pelo OJS. Na sequência, com os dados dessas revistas, realizou-se o acesso com a finalidade de verificar e descrever o ambiente de vocabulário controlado de periódicos científicos eletrônicos.

Na etapa 2, a realização da sistematização do conteúdo dos artigos do periódico científico eletrônico em formato de texto completo e sua inclusão em um banco de dados foi de grande relevância para o processo de compatibilização sistemática, considerando a pesquisa sistêmica das palavras-chave não compatibilizadas a partir dos padrões de Hearst. Essa etapa é considerada como atividades de planejamento, coleta e organização dos dados necessários para a execução das demais etapas. Vale acrescentar, que a coleta dos artigos foi realizada a partir da base de dados da VCPC *Tools*.

A identificação dos padrões de Hearst a partir das palavras-chave não compatibilizadas, objeto da etapa 3, tornou-se necessária para extrair do texto do artigo os excertos, que poderiam constar como possíveis padrões léxico-sintáticos a partir do termo chave. Na sequência, justifica-se a necessidade de localizar nesses excertos as palavras-chave dos artigos as quais não foram compatibilizadas pelo estudo de Santos (2015), com a finalidade de mapear a

sistemática das palavras-chave e compatibilizar com o vocabulário controlado de referência TBCI.

O mapeamento da sistemática das palavras-chave com o vocabulário controlado e a compatibilização, previsto na etapa 4, deu-se por meio dos padrões de Hearst (1992; 1998), considerando os níveis de correspondência de Neville (1970). Essa ação tornou-se uma contribuição significativa para a complementação da compatibilização de Santos (2015), e consequentemente essas compatibilizações enriquecem o vocabulário controlado do periódico por meio do estabelecimento de vínculos das palavras-chave dos artigos com os termos do vocabulário de referência (TBCI). Dessa forma, estabeleceu-se a interoperabilidade semântica.

O desenvolvimento do mapeamento reverso das palavras-chave compatibilizadas pelos processos de Santos (2015), a partir dos níveis de correspondência de Neville (1970), é atividade prevista na etapa 5, que somou para a instanciação do modelo de interoperabilidade semântica dos vocabulários controlados dos periódicos em questão. Por outro lado, foi possível estruturar, a partir da proposição do modelo de interoperabilidade semântica, a interoperabilidade sistêmica entre o protótipo da interface de busca e as bases de dados (TemaTres e VCPC *Tools*) por meio de API. As ações mapeadas no Quadro 15 e os métodos realizados para desenvolver as atividades de cada ação, serão descritas nos itens 5.1 ao 5.5.

5.1 Reconhecer o ambiente de vocabulário controlado de periódicos científicos eletrônicos gerenciados pelo OJS - Etapa 1

Ação 1: As buscas em bases de dados de teses e dissertações foram estabelecidas por meio dos termos que compõem a necessidade de busca: “vocabulário controlado” e “periódicos científicos”; “sistemas de organização do conhecimento” e “periódicos científicos”; “vocabulário controlado” e “revistas científicas”; "sistemas de organização do conhecimento" e "revistas científicas". A necessidade de busca é traduzida para recuperar, na BDTD, as teses e dissertações relacionadas com a temática e objeto deste estudo. A organização dos registros para análise foi realizada por meio de tabelas e planilhas eletrônicas.

Ação 2: Diante da busca por estudos relacionados, uma importante fonte é a base *Lattes* de currículos do CNPQ, de modo que se optou pela realização de buscas nos currículos *Lattes* de autores envolvidos com a temática. As buscas foram organizadas utilizando a opção do sistema de “Busca Avançada (por Assunto)”, no campo “todas essas palavras” nas bases “Doutores” e “Demais pesquisadores (Mestres, Graduados, Estudantes, Técnicos, etc.)”, nacionalidade brasileira e estrangeira, em todos os países. Como termo de busca, utilizou-se o

conjunto de palavras *vocabulário controlado periódicos científicos eletrônicos* e *controle vocabulário periódicos científicos eletrônicos*, caracterizando resultado 1 e 2. A síntese dos resultados foi organizada em planilhas eletrônicas de forma sistemática, realizando a eliminação de redundâncias. A partir dessa súmula, realizou-se, de maneira manual, a consulta dos currículos e análise das publicações de cada autor, sendo verificado o item produção bibliográfica (artigos completos publicados em periódicos; artigos aceitos para publicação; livros e capítulos; texto em jornal ou revista - magazine; trabalhos publicados em anais de eventos; apresentação de trabalho e palestra; partitura musical; tradução; prefácio, posfácio; outra produção bibliográfica).

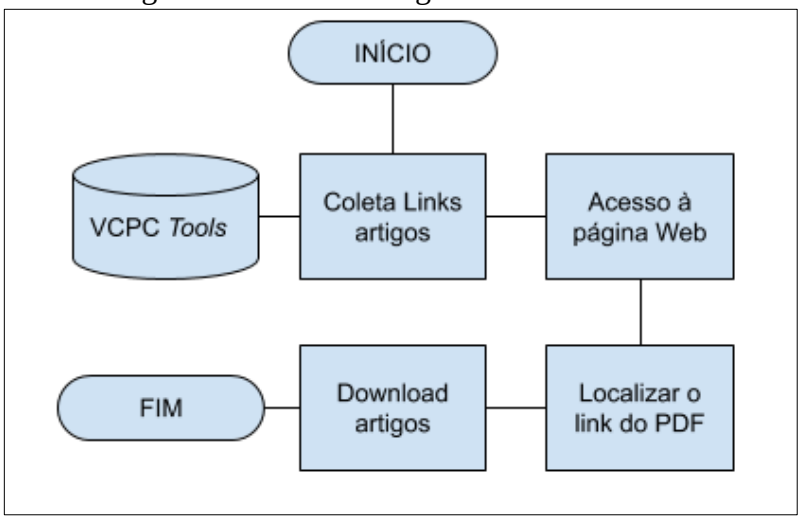
Ações 3 e 4: Foram realizadas visitas aos portais de periódicos com a finalidade de analisar o vocabulário controlado e sua disponibilização a partir da recuperação dos trabalhos localizados por meio dos resultados de análise da ação 2 - currículos *Lattes*. O recorte dos trabalhos foi realizado de maneira intelectual, com a análise do foco de cada estudo na aplicação do vocabulário controlado em periódicos científicos eletrônicos gerenciados pelo sistema OJS. Com base na execução e análise dos resultados dessa etapa, foi possível delinear o *corpus* de periódicos científicos eletrônicos em que foram aplicados os procedimentos das próximas etapas, como validação da proposta.

5.2 SISTEMATIZAR O CONTEÚDO DOS ARTIGOS DO PERIÓDICO CIENTÍFICO ELETRÔNICO EM FORMATO DE BUSCA EM TEXTO COMPLETO A PARTIR DA BASE DE DADOS DA VCPC *TOOLS* - ETAPA 2

Ação 1: As ações desta etapa tiveram como principal finalidade a construção de uma base de dados para inserir parte dos textos na íntegra dos artigos dos periódicos para execução dos procedimentos da etapa 3. O levantamento da coleção dos artigos foi definido a partir da execução das ações da etapa 1, que subsidiou a seleção do *corpus* que foi formado pelos artigos dos periódicos Informação & Informação, Informação@Profissões e Discursos Fotográficos, descritos nos resultados da etapa 1, item 6.1.

A coleta dos textos foi automatizada por meio de um algoritmo codificado em PHP que teve como principais funcionalidades: coletar todos os links dos artigos que estão armazenados nas bases de dados da VCPC *Tools*, aplicados nos periódicos; acessar o endereço do link, efetuar a varredura na página para localizar o endereço do arquivo em PDF; fazer o *download* dos artigos em arquivo. Na Figura 8 estão sistematizadas essas funcionalidades.

Figura 8 – Diagrama de Blocos - Algoritmo de coleta texto na íntegra

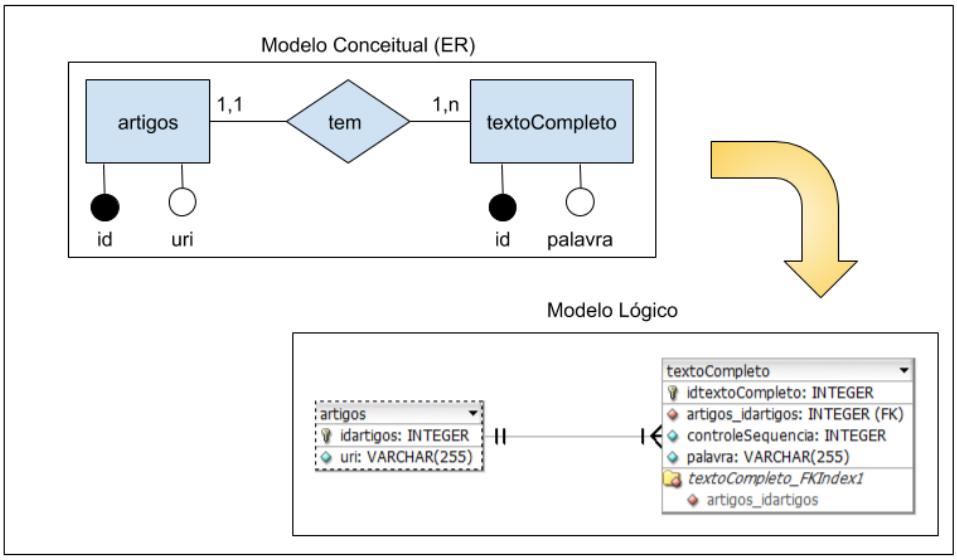


Fonte: Elaborado pelo autor (2020).

Ação 2: A conversão de PDF para TXT foi realizada por meio da ferramenta *Adobe Acrobat Reader* de maneira operacional.

Ação 3: A ação de desenvolvimento deste projeto conceitual e lógico do banco de dados, que recebeu esta coleta de dados de maneira automatizada, é, portanto, o planejamento de uma base de dados que passa por duas fases. Heuser (2004) define como projeto do banco de dados, cujas fases são: modelo conceitual e modelo lógico. O fluxo deste processo está ilustrado na Figura 9.

Figura 9 – Exemplo de modelos de banco de dados



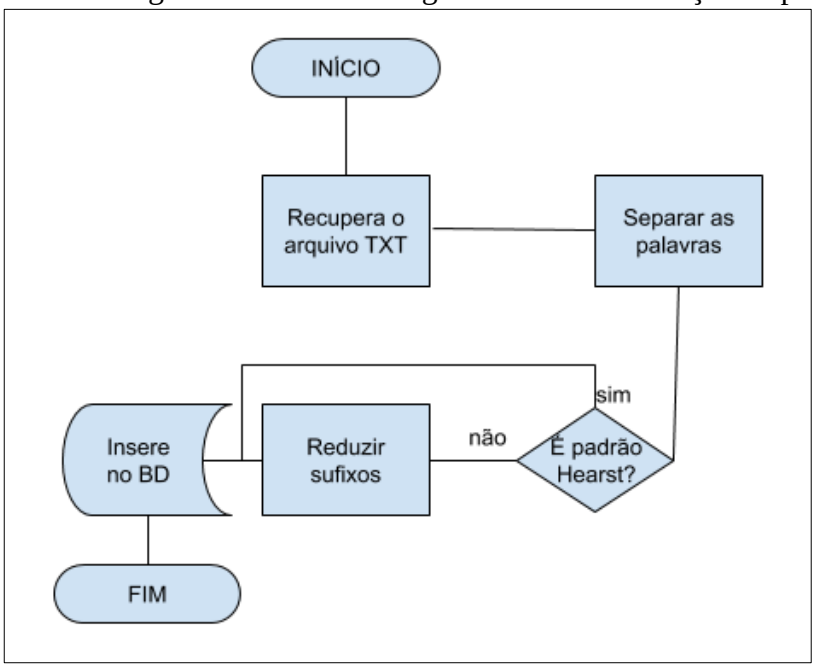
Fonte: Elaborado pelo autor (2020).

O modelo conceitual ou entidade-relacionamento (Figura 9) é o início do planejamento de um banco de dados, ou seja, neste processo identificam-se as entidades e os possíveis

atributos, além dos relacionamentos estabelecidos entre as entidades, nos quais constam as cardinalidades. As cardinalidades são definidas por meio de questões, por exemplo: “Um artigo pode ter quantos registros de texto completo? R: 1 ou N”, o que resultou na cardinalidade <1,n> do lado da entidade <textoCompleto> na Figura 9. Após essa modelagem, converte-se para o modelo lógico, o qual passa a ter uma descrição mais detalhada no que tange à criação em um sistema gerenciador de banco de dados, e cada entidade passa a ser chamada de tabela. Os atributos passam a ter tipo de valores que serão armazenados, os relacionamentos passam a constar com os atributos estrangeiros a eles referenciados.

Ação 4: A estruturação dos artigos em arquivos no formato TXT em linguagem SQL para inserção em banco de dados contou com a automatização por meio de um algoritmo que está mapeado na Figura 10, com as funcionalidades: recuperar os arquivos TXT, separar as palavras, verificar se é um padrão, retirar plurais e sufixos nos casos em que a palavra não seja o termo-chave considerado para a identificação do padrão.

Figura 10 – Diagrama de Blocos - Algoritmo de identificação de padrão



Fonte: Elaborado pelo autor (2020).

Na função <separar as palavras> (Figura 10), o texto do artigo será fragmentado em palavras, conforme apresentado no Quadro 16 a fragmentação do trecho “...*já no período que se segue o esforço foi desenvolver tecnologias que facilitassem as atividades da biblioteca, tais como: catalogação, indexação e serviços de circulação...*” (JESUS; CUNHA, 2019, p.5). O conjunto de palavras que foi armazenado no banco de dados, formou-se a partir do termo-chave

que identifica o padrão, sendo composto por 20 palavras que antecedem e 20 que sucedem, contabilizando 41 palavras que se tornaram registros no banco de dados.

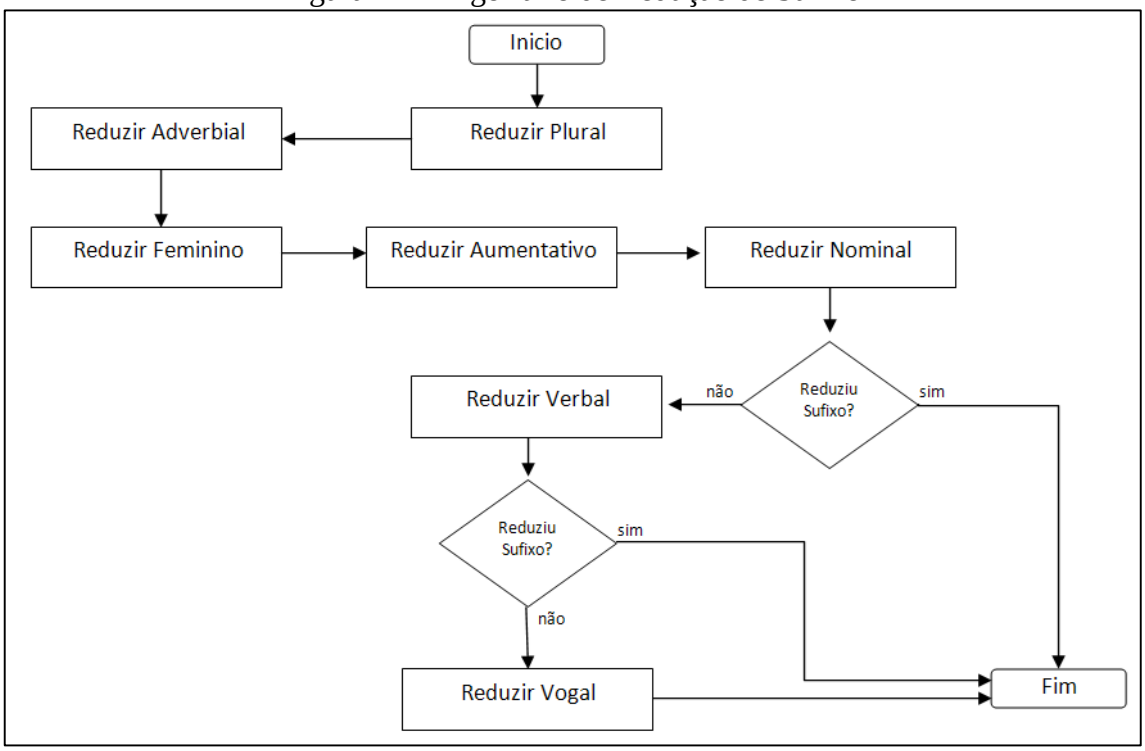
Quadro 16 – Exemplo de fragmentação de texto

SEQ	PALAVRA	SEQ	PALAVRA	SEQ	PALAVRA	SEQ	PALAVRA	SEQ	PALAVRA
1	já	6	segue	11	tecnologias	16	da	21	indexação
2	no	7	o	12	que	17	biblioteca,	22	e
3	período	8	esforço	13	facilitassem	18	tais	23	serviços
4	que	9	foi	14	as	19	como:	24	de
5	se	10	desenvolver	15	atividades	20	catalogação,	25	circulação

Fonte: Elaborado pelo autor (2020)

Na sequência 18 e 19 (Quadro 16) encontra-se um dos padrões de Hearst (1992, 1998). O algoritmo registrou, na base de dados, uma sinalização positiva desta sequência. No algoritmo de identificação de padrões (Figura 10), serão retirados os sufixos utilizando o algoritmo de Santos (2015, p.98), adaptado de Orengo, Buriol e Coelho (2007), apresentado na Figura 11. Esse procedimento foi executado para a sequência de palavras que antecedem e sucedem o termo-chave do padrão.

Figura 11 – Algoritmo de Redução de Sufixo



Fonte: Adaptado por Santos (2015, p.98) a partir de Orengo, Buriol e Coelho (2007).

Os termos-chave utilizados para marcar um padrão foram: tais como, tal ou tais, como, ou outro, ou outros, e outro, e outros, incluindo, especialmente, principalmente, particularmente, em especial, em particular, de maneira especial, sobretudo. Tomou-se a decisão estratégica de armazenar somente os possíveis padrões, com a finalidade de diminuir a quantidade de registros no banco de dados e superar a performance nas consultas.

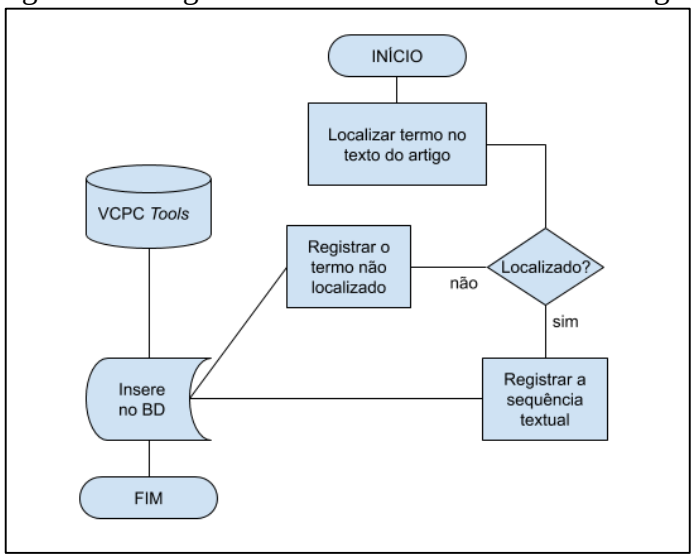
5.3 IDENTIFICAR OS PADRÕES DE HEARST A PARTIR DAS PALAVRAS-CHAVE NÃO COMPATIBILIZADAS - ETAPA 3

Nesta etapa, o *corpus* de tratamento e análise foram as palavras-chave não compatibilizadas de maneira automática nos processos de Santos (2015), que são: idênticas, idênticas retirando palavras vazias, idênticas retirando plurais e sufixo, índice contido de termo em palavra-chave.

Ação 1: A estruturação de consultas e a localização de cada padrão de Hearst (1992, 1998) foram realizadas de maneira automatizada. A localização dos padrões de Hearst (1992, 1998), considerando o estudo de Moreira, Santos e Vitorini (2017), foi sinalizada na base de dados por meio de um campo do tipo booleano, para melhor construção de resultados, conforme Figura 10. Para cada palavra-chave não compatibilizada foi realizada a consulta ao texto do artigo, logo constou o excerto para listar na planilha. Essa localização foi realizada utilizando os processos de Santos (2015) para compatibilização de palavras-chave (idênticas, idênticas retirando palavras vazias, idênticas retirando plurais e sufixo), exceto o índice contido de termo. As análises foram realizadas de maneira intelectual, com a utilização de planilhas. Por meio de padrões pré-estabelecidos no algoritmo da quantidade de sintagmas nominais que antecedem e sucedem (20 palavras), gerou-se uma planilha com as palavras-chave/termos com o respectivo padrão de Hearst (1992, 1998) e os demais sintagmas nominais, esse processo está ilustrado no algoritmo da Figura 12.

As palavras-chave não compatibilizadas estão representando um determinado artigo ou vários artigos. A busca pela localização das palavras-chave *a priori* foi realizada somente no artigo relacionado.

Figura 12 – Algoritmo de busca dos termos nos artigos



Fonte: Elaborado pelo autor (2020).

Ação 2: A análise de palavras anteriores e posteriores aos padrões de Hearst (1992, 1998), a partir de planilhas e relatórios gerados, forneceu suporte na tomada de decisão de compatibilização por meio da verificação no vocabulário de referência.

No Quadro 17 são apresentados os padrões de Hearst (1992, 1998) fundamentado no estudo de Moreira, Santos e Vitorini (2017).

Quadro 17 – Os padrões léxico-sintáticos aplicados na pesquisa

PADRÕES DE HEARST (1992, 1998)	**PADRÕES APLICADOS**
NP_0 such as NP_1 {,NP_2 ... , (and | or) NP_i}	(1) SN (tais como | como) SN { , SN ... , } (e | ou) SN
such NP_0 as {NP_1,}* {(and | or)} NP_2	(2) tal(is) SN como {(SN,)*(ou|e)} SN
NP_1 {, NP_1}* {,} or other NP_0	(3) SN {, SN}* {,} ou outro(s) SN
NP_1{, NP_2}* {,} and other NP_0	(4) SN {, SN}* {,} e outro(s) SN
NP_0 {,} including { NP_1,}* {or | and} NP_2	(5) SN {,} incluindo {SN,}*{ou|e} SN
NP_0 {,} especially { NP_1,}* {or | and} NP_2	(6a) SN {,} especialmente {SN,}*{ou|e} SN (6b) SN {,} principalmente {SN,}*{ou|e} SN (6c) SN {,} particularmente {SN,}*{ou|e} SN (6d) SN {,} em especial { SN,}*{ou|e} SN (6e) SN {,} em particular { SN,}*{ou|e} SN (6f) SN {,} de maneira especial { SN,}*{ou|e} SN (6g) SN {,} sobretudo { SN,}*{ou|e} SN

Fonte: Moreira, Santos e Vitorini (2017, p.175).

Ação 3: A organização dos resultados em planilha das análises, como ilustrado um exemplo no Quadro 19, seguiu a categorização de nove classes de análise (Quadro 18): N_CORRESP_P; N_CORRESP_FORTE; CORRESP_FRACO; CORRESP_FORTE;

CORRESP_LOC_REF; N_CORRESP_P_REF; N_CORRESP_FRACO; LIXO; e DUPLICADO.

Quadro 18 - Categorização dos excertos em classes de correspondência

Nº	Classe	Descrição
1	N_CORRESP_P	Corresponde com a palavra-chave e não existe correspondência com um padrão Hearst.
2	N_CORRESP_FORTE	Não corresponde com a palavra-chave, porém é um padrão com uma forte relação hierárquica.
3	CORRESP_FRACO	Corresponde com a palavra-chave, porém o padrão não gera uma relação hierárquica significativa.
4	CORRESP_FORTE	Corresponde com a palavra-chave e também com o padrão de maneira importante.
5	CORRESP_LOC_REF	Corresponde com a palavra-chave, porém a localização do excerto é nas referências bibliográficas.
6	N_CORRESP_P_REF	Não corresponde com a palavra-chave, nem com o padrão, e foi localizado nas referências bibliográficas.
7	N_CORRESP_FRACO	Não existe uma correspondência com a palavra-chave e remete a uma relação hierárquica fraca.
8	LIXO	Quando é localizado em partes do artigo que não são significativas, como cabeçalhos, filiações dos autores, acesso às referências, etc.
9	DUPLICADO	Quando um excerto aparece duas ou mais vezes na base de dados.

Fonte: Elaborado pelo autor (2020).

As classes listadas no Quadro 18 foram utilizadas no registro de exemplificação apresentado no Quadro 19 na coluna <Análise>, o qual recebeu a categorização.

Quadro 19 – Modelo de planilha de análise padrões

Palavra-Chave	Antecessor	Termo-chave	Sucessor	Análise
Análise Categoria	pesquisador propõe inferências e interpretações a propósito dos objetivos previamente determinados. Na Análise de Conteúdo podem ser utilizadas diferentes técnicas,	tal(is) como:	Análise Categorial; Analise de Avaliação; Análise da Enunciação; Análise Proposicional do Discurso e Análise de Expressão (BARDIN, 2009). No	CORRESP_FORTE

Fonte: Elaborado pelo autor (2020).

No Quadro 19 o termo-chave são os termos que representam os padrões de Hearst, a partir deles foram construídos os excertos com o antecessor e sucessor ao termo-chave. Esta finalização da análise com base no modelo do Quadro 19 deu o suporte para verificação junto ao vocabulário controlado de referência, que está sedimentado nas ações da etapa 4 (5.4).

5.4 MAPEAR, POR MEIO DOS PADRÕES DE HEARST (1992; 1998), A SISTEMÁTICA DAS PALAVRAS-CHAVE COM O VOCABULÁRIO CONTROLADO E COMPATIBILIZAR COM OS NÍVEIS DE CORRESPONDÊNCIA - ETAPA 4

Com base nos resultados gerados na etapa 3, passa-se para a análise semiautomática, por meio da qual se realizou a compatibilização da palavra-chave com o termo do vocabulário controlado e a convalidação da sistemática, localizada para o termo do vocabulário controlado.

Ação 1: O registro do relacionamento da palavra-chave ocorreu por meio do mapeamento e compatibilização da palavra-chave, do TBCI e do vocabulário Comunicação Visual. No caso da localização da sistemática para os termos do vocabulário controlado do periódico, serão validadas as relações hierárquicas no TBCI. Todos os procedimentos serão registrados na tabela/planilha de resultados.

Após a tomada de decisão por registrar a sistemática localizada, passa-se aos procedimentos de inclusão do termo e dos relacionamentos hierárquicos no vocabulário do periódico. Esse processo requer, da validação junto ao TBCI, a verificação da interoperabilidade deste termo, que já foi realizada na atividade anterior. Aqui ocorre somente o registro no vocabulário controlado, por meio da instanciação de vínculos com vocabulário de referência TBCI. Para todas essas ações que culminam no mapeamento das palavras-chave foi utilizada a proposta de Neville(1970) em relação à matriz de compatibilização para o registro (Quadro 20).

Quadro 20 – Modelo de mapeamento de palavras-chave com padrões de Hearst

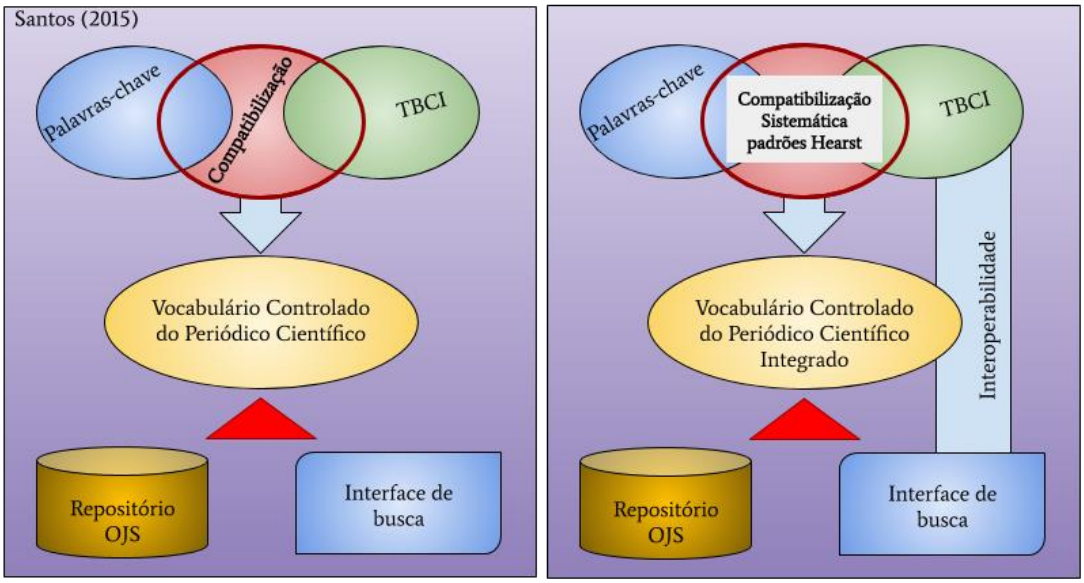
Palavra-Chave	**Termo TBCI**	**Níveis de correspondência (Neville, 1970)**	**Mapeamento**
Revisão sistemática de literatura	Modelos de Análise	Caso 4	CORRESP_FORTE

Fonte: Elaborado pelo autor (2020).

Com a execução dessa etapa, os processos foram sintetizados e podem ser visualizados na Figura 13. Apresenta-se a sequência do estudo dividido em duas partes complementares: a)

proposta de Santos (2015), origem dos estudos, que realizou o controle de vocabulário por meio da compatibilização das palavras-chave; b) complementação do processo de compatibilização com a compatibilização sistemática realizada de maneira semiautomática por meio dos padrões de Hearst (1992, 1998), bem como a proposta de interoperabilidade semântica entre os vocabulários controlados dos periódicos que compuseram o estudo de caso e o vocabulário de referência, culminando no protótipo da interface de busca.

Figura 13 – Modelo de interoperabilidade



Fonte: Elaborado pelo autor (2020).

O vocabulário controlado do periódico científico integrado com a inclusão das estruturas hierárquicas validadas passou a fazer parte do modelo de interoperabilidade entre os vocabulários controlados e interface de busca. A proposta de Santos (2015) ilustrada na Figura 13 foi alvo das ações da etapa 5, estruturada no item 5.5, onde se desenvolveu o mapeamento reverso. O modelo da Figura 13 foi completado a partir da descrição da próxima etapa, em que constam o mapeamento das palavras-chave já compatibilizada por Santos (2015) e de modo geral a implantação experimental do modelo de interoperabilidade.

5.5 Desenvolver o mapeamento reverso das palavras-chave compatibilizadas pelos processos de Santos (2015) a partir dos níveis de correspondência - Etapa 5

Esta etapa está relacionada com o desenvolvimento do mapeamento e implantação experimental do modelo de interoperabilidade, perpassando pelas ações planejadas para o mapeamento reverso das palavras-chave. Além do tratamento dos procedimentos internos da

etapa 5, são sistematizadas as ações de maneira a sintetizar os processos realizados nas demais etapas da proposta teórico-metodológica.

Ação 1: A modelagem e execução do mapeamento reverso de palavras-chave compatibilizadas por Santos (2015) foi realizada observando a proposta de Neville (1970), na qual se utilizam os níveis de correspondência da matriz de compatibilização (ver o modelo de mapeamento no Quadro 20). Seguiu-se este modelo observando os casos de correspondência a partir dos processos de compatibilidade das palavras-chave. A listagem de palavras-chave foi coletada da ferramenta VCPC *Tools*, com os respectivos mapeamentos de compatibilização.

Ação 2: O desenvolvimento do modelo conceitual e lógico de interoperabilidade passa por um planejamento metodológico e de gestão da estrutura. Listam-se os vocabulários controlados a serem interoperados, com base nos fundamentos teóricos e técnicos. Conta-se com atividades de desenvolvimento de uma macroestrutura do VCPC-CI, por meio da classificação das áreas e subáreas de avaliação da Capes. Nesse caso, a área de avaliação *Comunicação e Informação*, que deu origem ao nome do vocabulário controlado (VCPC-CI), e subordinadas a ela estão as subáreas: Ciência da informação - CI e Comunicação – CO, as quais deram origem aos dois metatermos mais gerais da macroestrutura.

Os procedimentos de preparação do TemaTres – *software* gestor do vocabulário controlado foram: sua instalação, configuração, criação de usuários para gerenciar o VCPC-CI, em um simulador de serviço *Web* - local. O requisito necessário para executar o servidor local é o Apache (EasyPHP – Devserver 17). Na estruturação do VCPC-CI no TemaTres contou-se, primeiramente, com o registro manual da macroestrutura, na sequência, a importação do Plano de Classificação Geral do TBCI e a importação do Plano Geral da Comunicação Visual, posteriormente, a definição de cada termo da macroestrutura como metatermo.

Ação 3: A organização dos termos para importação no vocabulário controlado integrativo dos periódicos - VCPC-CI dependeu da estruturação de uma consulta em SQL que foi executada no banco de dados da VCPC *Tools* de cada revista do *corpus*, a partir da ferramenta *MySQL Admin* para exportação dos termos em suporte de arquivo TXT. De posse dos termos em TXT, passa-se à estruturação no formato de importação no TemaTres (Figura 14). Neste caso, foi utilizado o formato texto etiquetado no suporte de arquivo TXT, utilizado por Santos, Fujita e Moreira (2018) para importação de registros de autoridades em formato MARC21 no suporte de arquivo XML no TemaTres.

Figura 14 – Formato de importação TemaTres - TXT etiquetado

```
Placas (Engenharia)
UF: Disks (Mechanics)
UF: Panels
UF: Structural plates
BT:  Placas e cascas elásticas
BT:  Análise estrutural (Engenharia)
RT: Cascas (Engenharia)
RT: Plates (Engineering)
NA: (UNAU000201540) Usado para trabalhos sobre placas como estruturas de engenharia. Para trabalhos sobre cascas

Trajes História
UF: Indumentária medieval
UF: Trajes medievais
UF: Costume, Medieval
RT: Costume History
NA: (UNAU000201549)

Contos espíritas
BT:  Ficção espírita
RT: Spiritual short stories
NA: (UNAU000201603)

Tabebuia caraiba
UF: Caraibeira
UF: Caroba-do-campo
UF: Cinco-em-rama
UF: Craibeira
UF: Ipê-amarelo-do-cerrado
UF: Para-tudo
```

Fonte: Santos, Fujita e Moreira (2018).

Por fim, importar os termos do VCPC dos periódicos: Informação & Informação e Informação@Profissões. O VCPC do periódico Discursos Fotográficos já estava na estrutura do TemaTres, portanto foi realizada uma requisição de exportação do vocabulário Comunicação Visual para importar no VCPC-CI. Ressalta-se que em todos os métodos de importação foi realizada a inserção das classificações dos termos gerais na macroestrutura.

Ação 4: A atividade de registro da interoperabilidade dos termos com o vocabulário de referência TBCI *a priori* foi realizada de maneira intelectual, devido ao fato de os testes de importação no TemaTres por meio do SKOS com conceitos vinculados a recursos externos não terem sidos satisfatórios na versão 3.0.

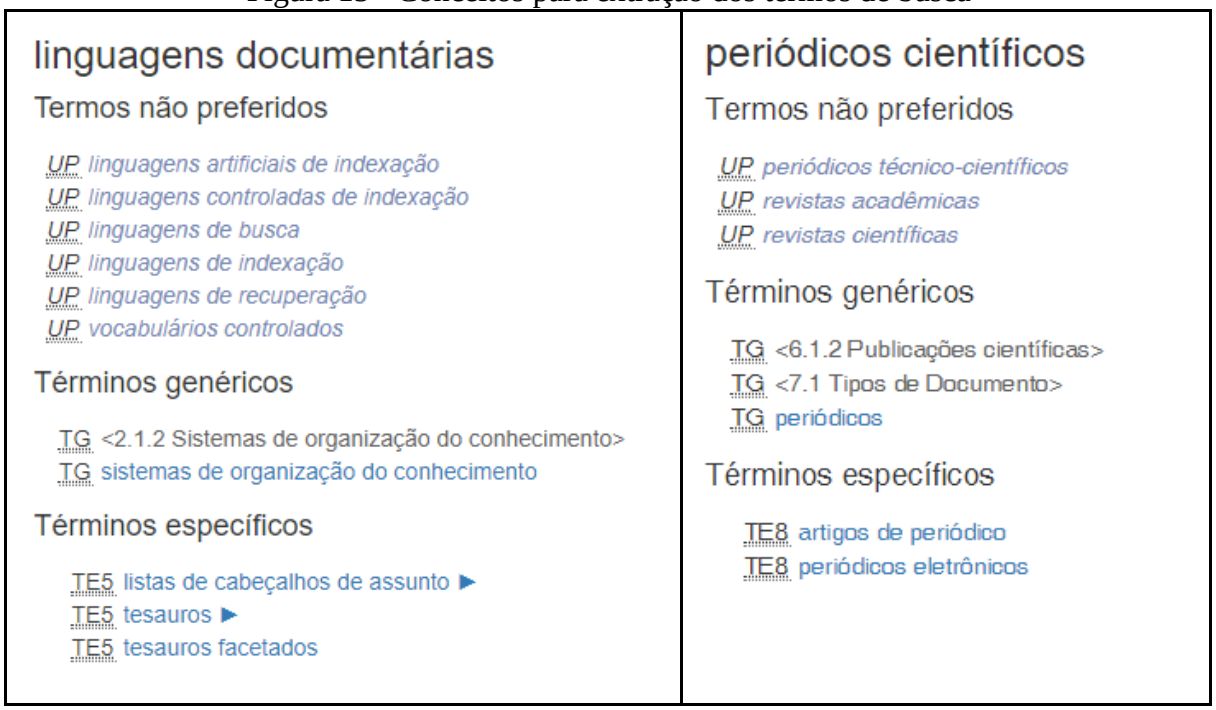
6 ESTUDO DE CASO PORTAL DE PERIÓDICOS DA COMUNICAÇÃO E INFORMAÇÃO DA UEL

Considerando-se os objetivos deste estudo, estão sistematicamente organizados os resultados da execução das cinco etapas de pesquisa deles decorrentes e sequencialmente apresentados nos itens 6.1, 6.2, 6.3, 6.4 e 6.5.

6.1 RECONHECER O AMBIENTE DE VOCABULÁRIO CONTROLADO DE PERIÓDICOS CIENTÍFICOS ELETRÔNICOS - OJS - ETAPA 1

Na ação 1 - Buscas em bases de dados de teses e dissertações, o primeiro passo foi o planejamento das estratégias de buscas, que se deu a partir do TBCI, com todos os termos que compõem a estrutura hierárquica do conceito "vocabulário controlado" e "periódico científico" (Figura 15). Especificamente no caso do "periódico científico", foi incluído um elenco rol dos termos relacionados considerado importante na representação deste conceito. A base de dados utilizada é a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD).

Figura 15 - Conceitos para extração dos termos de busca



Fonte: TBCI (2014).

A estratégia de busca foi estruturada com operadores *OR* e *AND*, concretizada da seguinte forma: *("periódicos científicos" OR "periódicos técnico-científicos" OR "revistas acadêmicas" OR "revistas científicas" OR "Publicações científicas" OR "Tipos de Documento" OR "periódicos" OR "artigos de periódico" OR "periódicos eletrônicos" OR "comunicação científica") AND ("linguagens documentárias" OR "linguagens artificiais de indexação" OR "linguagens controladas de indexação" OR "linguagens de busca" OR "linguagens de indexação" OR "linguagens de recuperação" OR "vocabulários controlados" OR "sistemas de organização do conhecimento" OR "listas de cabeçalhos de assunto" OR "tesauros" OR "tesauros facetados").*

Foram recuperados 30 registros, dentre os quais dois são duplicados. A partir da leitura e análise dos resumos, constataram-se 9 registros relacionados indiretamente com a temática e somente uma dissertação relacionada diretamente com este estudo. Os registros resultados da busca são apresentados no Quadro 21.

Quadro 21 - Pesquisa de trabalhos relacionados com a temática desta tese – vocabulários controlados em periódicos científicos eletrônicos

Título	Autor/ano	Instituição	Documento
Vocabulário controlado em periódicos científicos eletrônicos: uma proposta de controle de termos	**Santos (2015)**	**UEL**	**Dissertação**
Analisando conteúdos e mapeando informação em periódicos eletrônicos: um estudo do periódico secundário PBCIB	Souza (2011)	UFPB	Dissertação
A arte de indexar artigos de periódicos: a política de indexação da seção de periódicos da Biblioteca Central da UFPB	Galvino (2012)	UFPB	Dissertação
A organização temática da informação em periódicos científicos eletrônicos: atribuição de palavras-chave na biblioteconomia e ciência da informação	Dias (2012)	UEL	Dissertação
Ambiente para geração e manutenção semiautomática de tesauros	Moreira (2005)	UFMG	Tese
Descritores em ciências da saúde na área específica da fonoaudiologia brasileira	Ostiz (2010)	USP	Tese

Continua

Continuação

Metodologia para seleção de termos equivalentes e descritores de tesauros: um estudo no âmbito do Direito do Trabalho e do Direito Previdenciário	Laipelt (2015)	UFRGS	Tese
Indexação automática por atribuição de artigos científicos da área de ciência da informação	Bandim (2017)	UFPE	Dissertação
Contribuição para a terminologia do processo de inteligência competitiva: estudo teórico e metodológico	Cervantes (2004)	UNESP	Dissertação
Terminologia e documentação: um estudo terminográfico sobre performance musical	Carvalho (2013)	UFMG	Dissertação

Fonte: Dados da pesquisa.

A ação 2 - Buscas na base *Lattes* de currículos de autores com a temática, resultou em dois conjuntos de registros, a partir da identificação dos trabalhos publicados sobre a disseminação dos vocabulários controlados das revistas na base de Currículo *Lattes*. No primeiro conjunto, foram localizados 32 registros (estratégia de busca: "vocabulário controlado periódicos científicos eletrônicos"). No segundo conjunto, foram localizados 26 registros (estratégia de busca: "controle vocabulário periódicos científicos eletrônicos"). Nas estratégias foram utilizadas "todas as palavras" que têm correspondência com o operador booleano OR.

Após a estruturação dos resultados em tabelas e a verificação de cada registro e eliminação de 18 registros duplicados de autores, passou-se à consulta dos currículos, a fim de localizar trabalhos que abordam a temática. Neste caso, utilizou-se do trabalho seminal de Santos (2015) como exemplar específico do assunto, localizado na BDTD. Do total de 40 autores, cinco têm trabalhos publicados com a temática. A materialização dos resultados está apresentada no Quadro 22.

Quadro 22 – Trabalhos Publicados Relacionados com Vocabulário Controlado em Periódicos Científicos - UEL

Trabalhos	Resumo	Tipo de publicação
Santos e Cervantes (2014)	Apresenta uma pesquisa relacionada com controle de vocabulário em periódicos científicos eletrônicos, coleta e análise do comportamento das palavras-chave da revista Informação & Informação, a partir do recorte temporal do ano de 2013.	Anais
Santos e Cervantes (2015a)	Aborda, nesta pesquisa, com mais ênfase nos processos de compatibilização das palavras-chave, por meio da ferramenta VCPC *Tools*, denominada VCI&I; os resultados são apresentados parcialmente.	Cap. Livro
Santos e Cervantes (2015b)	Apresenta a proposta teórico-metodológica consolidada e implantada com as demonstrações dos resultados finais.	Anais
Santos, Cervantes, Londero e Goncalez (2016)	Inicia a análise para executar o projeto de implantação em outra área do conhecimento, periódico Discursos Fotográficos da área de Comunicação fundamentado em Santos e Cervantes (2015b).	Anais
Rodrigues, Pereira, Londero e Cervantes (2017)	Analisa as palavras-chave atribuídas pelos autores para torná-las descritores, trabalha os aspectos cognitivos com especialista da área para a composição do domínio e subdomínio.	Anais
Santos, Cervantes e Londero (2018)	Trabalha a concretização da proposta metodológica da construção do vocabulário controlado do periódico Discursos Fotográficos com base na metodologia de coleta e análise de palavras-chave de Santos e Cervantes (2015b).	Cap. Livro
Total de Publicações	**-**	**6**

Fonte: Dados da pesquisa.

A seleção dos periódicos está vinculada à produção dos autores conforme registrada na base *Lattes* de currículos. No Quadro 22, linha 5, o coautor "Pereira" foi recuperado somente a partir da recuperação do trabalho, ele não constou nos resultados da busca na base *Lattes* de currículos. Em análise das publicações descritas no Quadro 22, a revista Informação & Informação é a primeira revista constante em uma publicação tratando do desenvolvimento e implantação do controle de vocabulário. A revista Discursos Fotográficos é a segunda revista trabalhada em projetos de implantação do controle de vocabulário. Na segunda etapa dos resultados frente à observação aos portais das revistas, verificou-se que a revista Informação &

Informação disponibiliza a interface de busca da VCPC *Tools*, por meio do índice de termos abordado por Santos e Cervantes (2014), Santos e Cervantes (2015a) e Santos e Cervantes (2015b).

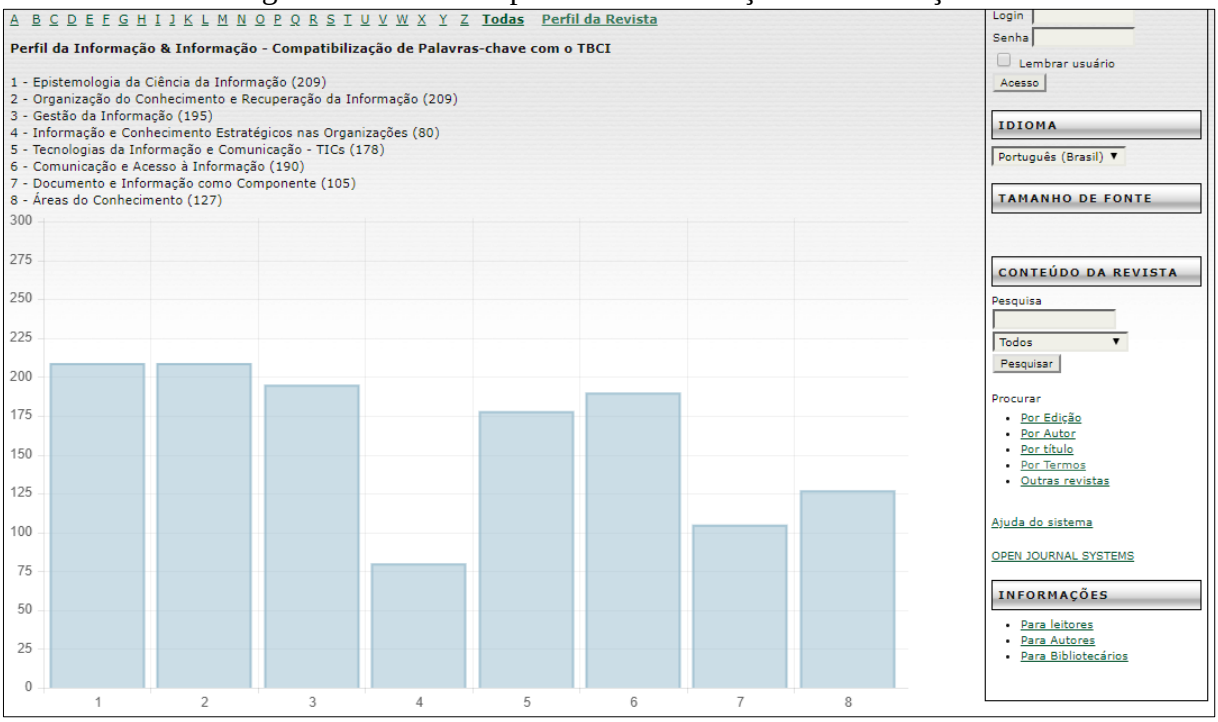
Os resultados das ações 3 e 4 são integrados para melhor organização. Com base nos trabalhos recuperados, listam-se os periódicos científicos eletrônicos observados: Informação & Informação e Discursos Fotográficos. No entanto, em consulta ao portal de periódicos da Universidade Estadual de Londrina (UEL), percebe-se que o Centro de Educação, Comunicação e Artes (CECA), por meio do Departamento de Comunicação e do Departamento de Ciência da Informação, conta com três periódicos, sendo dois na área da CI e um em Comunicação. Optou-se por incluir a observação do periódico Informação@Profissões. A partir da localização dos portais gerenciados pelo OJS desses periódicos científicos, realizou-se o acesso a cada um, respectivamente, para a observação. Na Figura 16, apresenta-se a interface com o índice de termos.

Figura 16 – Interface de busca por meio do índice de termos do Informação & Informação

Fonte: Portal do periódico Informação & Informação (2020).

Observa-se, na interface de busca da VCPC *Tools*, uma lista de termos e uma numeração na sequência. Esse número é a quantificação de artigos que estão vinculados com o termo (Figura 16). Ao clicar no termo, são visualizadas as relações daquele termo, nesse caso, a referência é o Tesauro Brasileiro em Ciência da Informação (TBCI), e ao clicar no número de artigos, são listados os títulos e o link para seu acesso. O perfil do periódico pode ser acessado pelo link <perfil da revista>, apresentado uma cópia na Figura 17.

Figura 17 – Perfil do periódico Informação & Informação



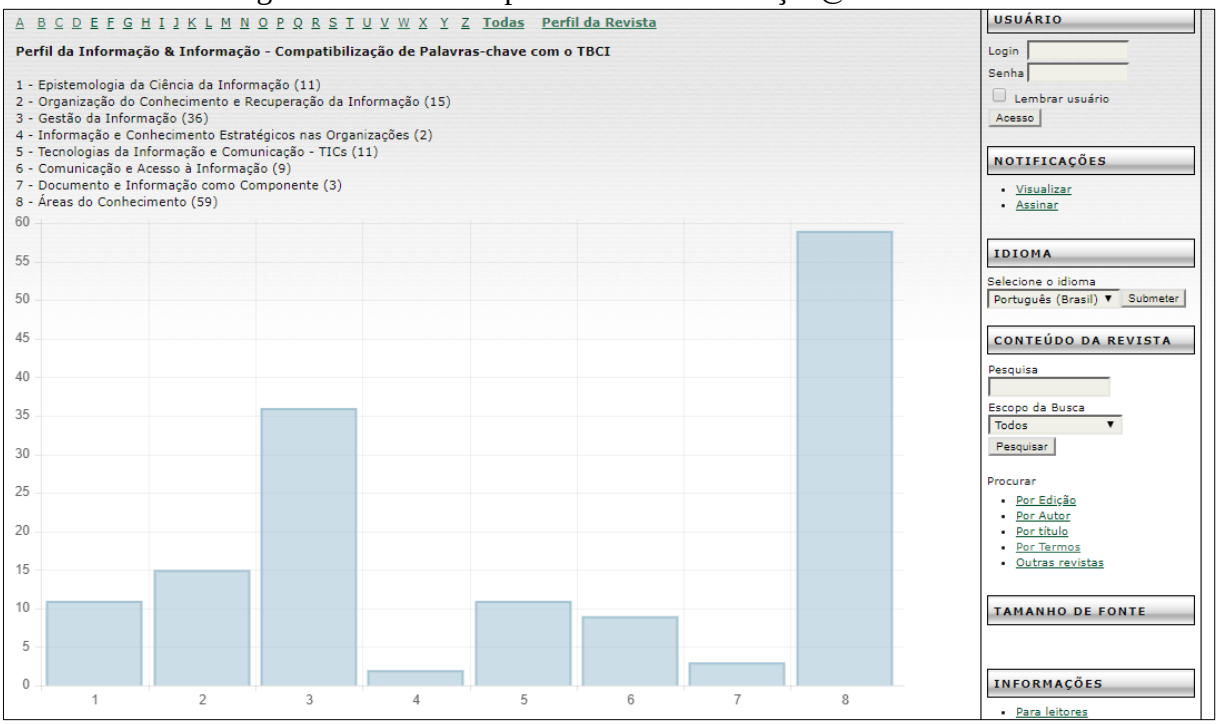
Fonte: Portal do periódico Informação & Informação (2020).

Em relação à segunda revista, Informação@Profissões, da área da CI, observa-se que consta a implantação de forma semelhante à que ocorreu na Informação & Informação, ilustrada a interface de busca por meio do índice e o perfil nas Figura 18 e Figura 19, respectivamente.

Figura 18 – Interface de busca por meio do índice de termos do Informação@Profissões

Fonte: Portal do periódico Informação@Profissões (2020).

Figura 19 – Perfil do periódico da Informação@Profissões



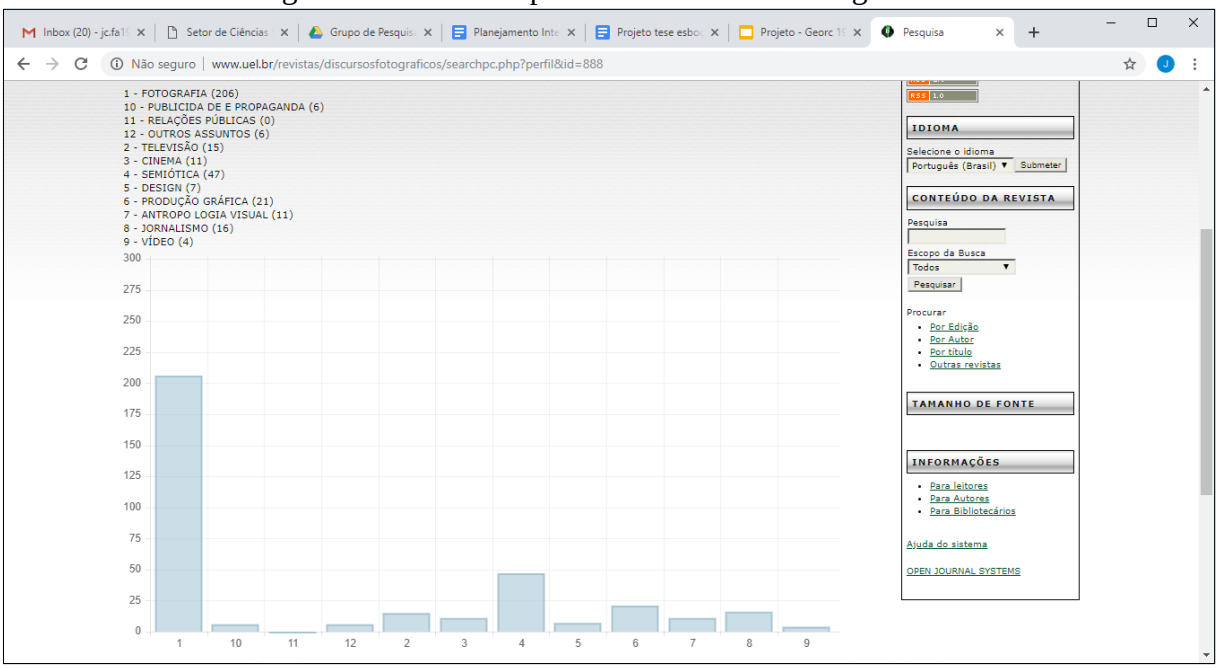
Fonte: Portal do periódico Informação@Profissões (2020).

No periódico Discursos Fotográficos é disponibilizado o índice de termos por meio da VCPC *Tools*. Os termos também são disponibilizados no TemaTres para gerenciamento do vocabulário controlado do periódico. Na Figura 20 apresenta-se o índice de termos, na Figura 21 o perfil e na Figura 22 o vocabulário controlado no TemaTres do periódico Discursos Fotográficos.

Figura 20 – Interface de busca por meio do índice de termos do Discursos Fotográficos

Fonte: Portal do periódico Discursos Fotográficos (2020).

Figura 21 – Perfil do periódico Discursos Fotográficos



Fonte: Portal do periódico Discursos Fotográficos (2020).

Cabe ressaltar que a busca dos artigos somente é possível a partir da interface da VCPC *Tools*: ao clicar no termo, são listados os artigos relacionados em ambas as revistas. O diferencial no modelo de apresentação da revista Discursos Fotográficos é a utilização do TemaTres para a gestão dos termos (Figura 22).

Figura 22 – Interface do TemaTres do vocabulário controlado do Discursos Fotográficos

Fonte: Portal do periódico Discursos Fotográficos (2020).

Os resultados dessa etapa devem ser compreendidos como desenvolvimento preliminar e de essência para o estabelecimento do *corpus* de análise no estudo piloto. Percebe-se a real necessidade do modelo conceitual de interoperabilidade entre os vocabulários

controlados dos periódicos científicos eletrônicos, definidos a partir deste ponto como *corpus*: Informação & Informação, qualificada como A2 no índice Qualis de avaliação de periódicos mantido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Informação@Profissões, qualificada como B1; e Discursos Fotográficos, qualificada como B1. Os periódicos estão vinculados à UEL, área de Comunicação e Informação dos departamentos de Comunicação e Ciência da Informação.

6.2 ESTRUTURAR O CONTEÚDO DOS ARTIGOS DO PERIÓDICO CIENTÍFICO ELETRÔNICO EM FORMATO DE BUSCA EM TEXTO COMPLETO - ETAPA 2

A organização do *corpus* nesta etapa resultou 1103 artigos científicos, sendo 333 da revista Discursos Fotográficos, 199 da revista Informação@Profissões e 571 da revista Informação & Informação. Os *downloads* dos arquivos em formato PDF foram executados pelos algoritmos: *downloads* dos artigos e localizador do link do PDF. Parte deles são apresentados como representação dos resultados na Figura 23 e Figura 24.

Figura 23 – Codificação algoritmo de download dos artigos

```
8  <?php
9      set_time_limit(0);
10     include_once ('conect.php');
11     include_once ('fun.listaartigos.php'); //lista os artigos das revistas
12     include_once ('fun.downloadpdf.php'); // copia os artigos em pdf
13     include_once ('fun.separapalavra.php');
14     include_once ('fun.loclinkpdf.php'); //localiza o link de download do artigo
15
16
17     $artigos= listaartigos($conecta_vcpc);
18     $i=0;
19     while ($i< count($artigos)){
20         $linkpdf= loclinkpdf($artigos[$i]['uri']);
21         echo 'Texto baixado: '. $linkpdf.'<br>';
22         downloadpdf ($linkpdf, "textos/".$artigos[$i]['id']);
23         $i++;
24     }
25
26     $i=0;
```

Fonte: Elaborado pelo autor (2020).

Na linha 20 da Figura 23 faz-se a chamada para a função de localização do link, resultante na Figura 24, linha 3. A varredura do link de cada artigo se deve pela não padronização da localidade do arquivo PDF no sistema OJS. Essa falta de padronização é decorrente dos diversos tipos de PDF suportados na plataforma e também dos aspectos de

editoração seguidos a cada edição publicada. O retorno da execução da codificação do algoritmo é o link do arquivo PDF – Figura 24.

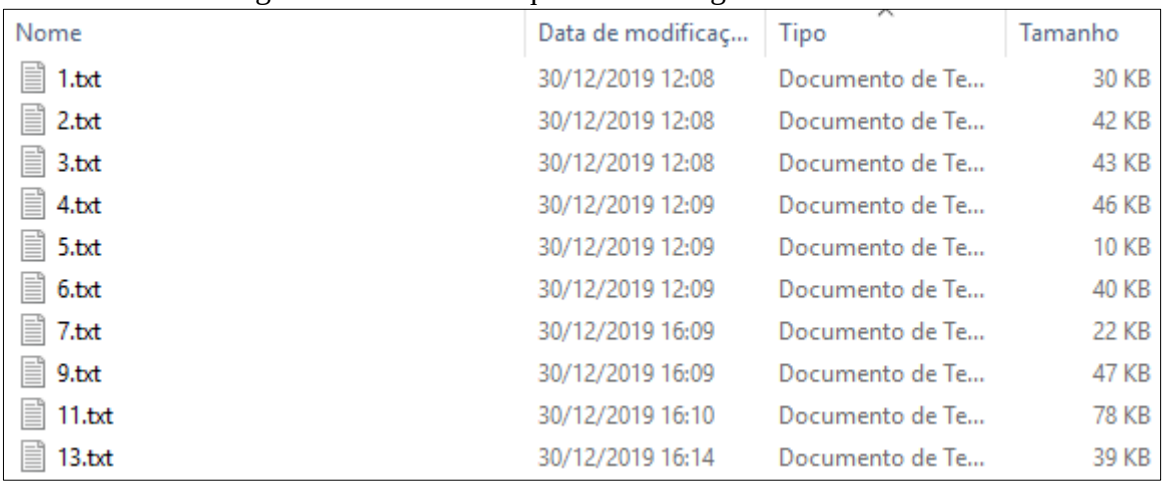
Figura 24 – Algoritmo de localização dos links que apontam para os arquivos PDF

```
3  function loclinkpdf ($link){
4      $html_pos= file_get_contents($link);
5      $html = str_split($html_pos);
6      //localiza o início do '>PDF</a>'
7      $i= stripos($html_pos, '>PDF</a>');
8      $partelink= substr($html_pos, $i-150, 150);
9      $i= stripos ($partelink, $link);
10     $partelink= substr ($partelink, $i);
11     $partelink_a = str_split($partelink);
12     $t= array_search ('"', $partelink_a);
13
14     //<a href="http://www.uel.br/revistas/uel/in
15
16     return (substr ($partelink, 0, $t));
```

Fonte: Elaborado pelo autor (2020).

Os resultados obtidos a partir da execução da ação 2, conversão em TXT, são visualizados de forma representativa por meio da *Figura 25*, a qual apresenta um *print* da interface de exploração de arquivos. Optou-se pela maneira mais confiável, a operacional, a partir do leitor nativo de PDF, o *Adobe Acrobat Reader*, devido aos resultados insatisfatórios dos diversos testes realizados com os conversores automatizados, entre eles *PDFtoText Class*, *PDF Parser* e *PDF2Text*. Um dos quesitos que dificultou a conversão é a codificação do arquivo, onde os caracteres especiais e acentuados não são reconhecidos, causando problemas na interpretação do conteúdo. Outra questão está relacionada aos diversos tipos de PDF que foram recuperados.

Figura 25 - Lista de arquivos dos artigos em formato TXT

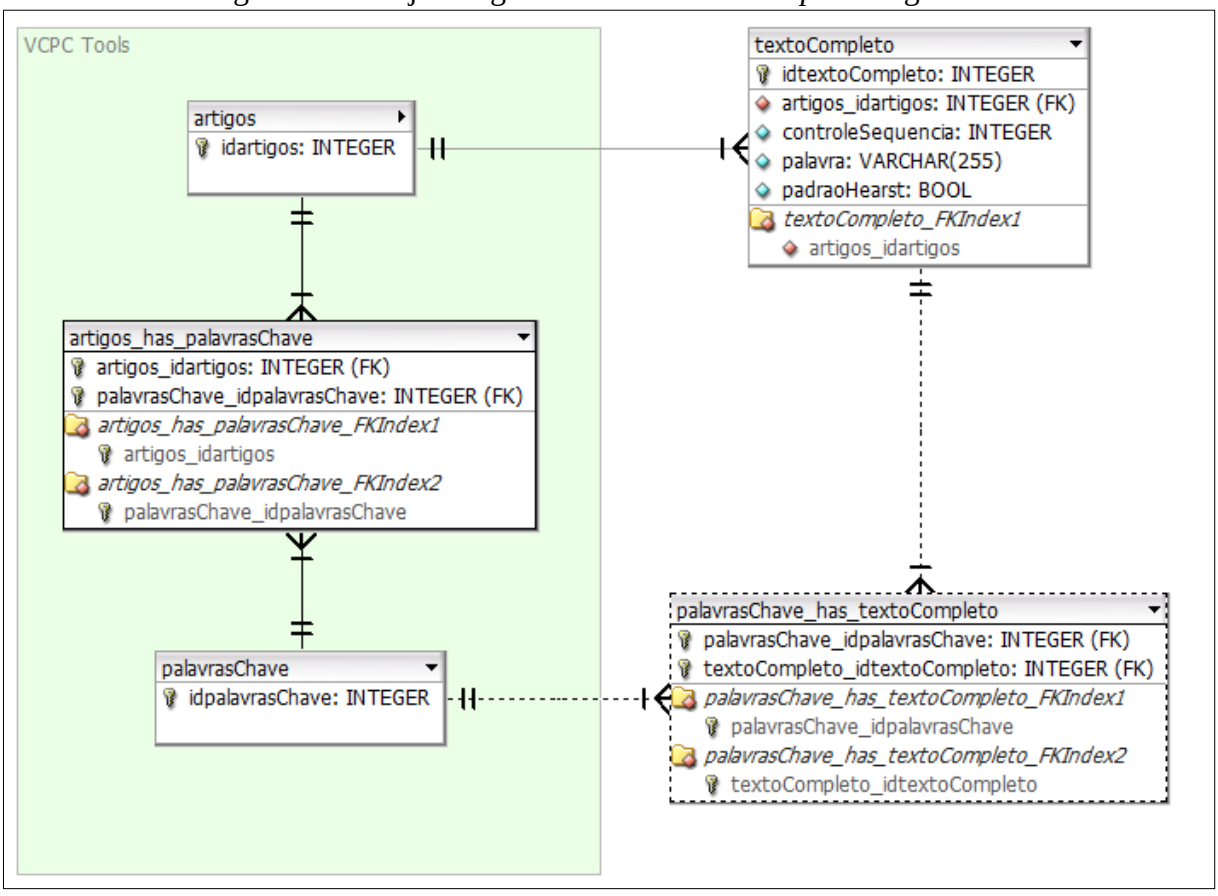


Nome	Data de modificaç...	Tipo	Tamanho
1.txt	30/12/2019 12:08	Documento de Te...	30 KB
2.txt	30/12/2019 12:08	Documento de Te...	42 KB
3.txt	30/12/2019 12:08	Documento de Te...	43 KB
4.txt	30/12/2019 12:09	Documento de Te...	46 KB
5.txt	30/12/2019 12:09	Documento de Te...	10 KB
6.txt	30/12/2019 12:09	Documento de Te...	40 KB
7.txt	30/12/2019 16:09	Documento de Te...	22 KB
9.txt	30/12/2019 16:09	Documento de Te...	47 KB
11.txt	30/12/2019 16:10	Documento de Te...	78 KB
13.txt	30/12/2019 16:14	Documento de Te...	39 KB

Fonte: Dados da pesquisa.

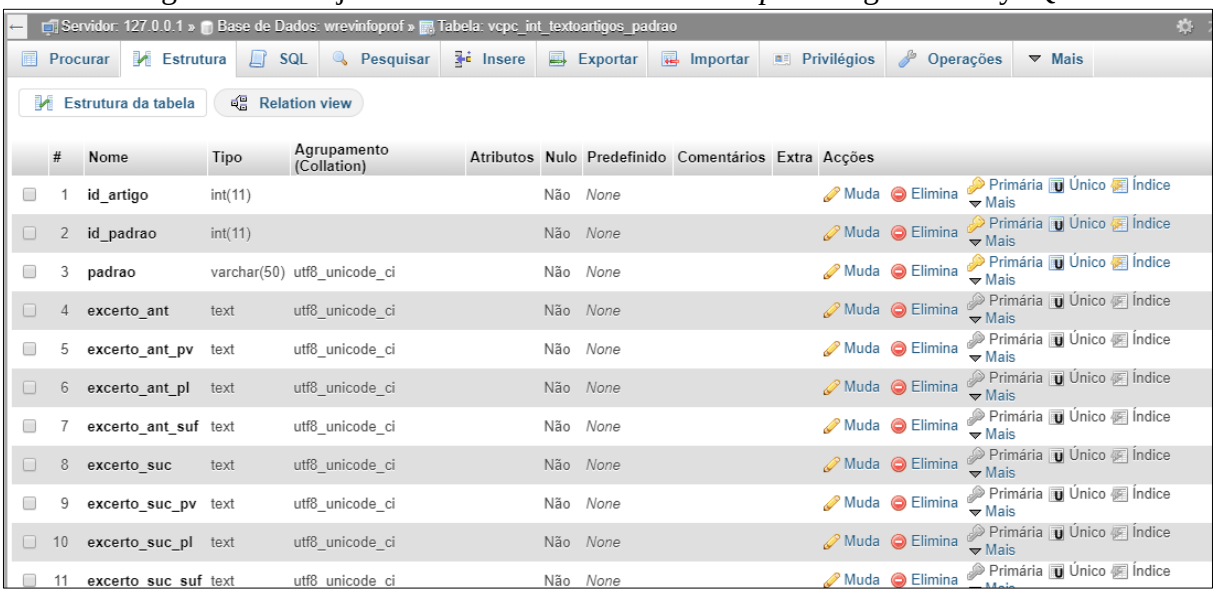
Na ação 3, que se refere ao desenvolvimento do projeto conceitual lógico do banco de dados, apresenta-se a modelagem do banco de dados, reproduzida na Figura 26, e apresenta-se a instanciação do banco de dados na base de dados de cada revista para recepção dos dados no SGBD MySQL, na Figura 27.

Figura 26 - Projeto lógico banco de dados *corpora* linguística



Fonte: Elaborado pelo autor (2020).

Figura 27 - Projeto físico do banco de dados *corpora* linguística MySQL



Fonte: Elaborado pelo autor (2020).

Na tabela <textoartigos> foram estabelecidos os vínculos com o artigo e a organização do *corpora* linguístico de cada artigo científico eletrônico. A verificação do termo-chave que identifica os padrões de Hearst (1992, 1998) foi marcada no campo <padraoHearst> em formato booleano (*true/false*). A tabela <palavraChave_has_textoartigos> armazena o vínculo da palavra-chave com a sua localização no texto do artigo científico.

Para efeitos de performance criou-se também a tabela <textoartigos_padrao>, onde foram armazenados os excertos compilados a partir dos termos-chave que identificam os padrões linguísticos. Após a inserção de todos os registros dos textos dos artigos, que em média têm aproximadamente 16 mil palavras, a tabela <textoartigos> totalizou mais de 1 milhão de registros para cada revista. Portanto, optou-se em trabalhar com a estrutura da tabela <textoartigos_padrao> e ainda, para efeitos desta tese, com um recorte dos dois últimos anos de cada revista, 2018 e 2019. A automatização desse processo está codificada por meio do algoritmo apresentado na Figura 28, representando como resultado parte da codificação.

Figura 28 - Codificação do algoritmo de identificação dos padrões

```
48    $artigos= listaartigos($conecta_vcpc);
49    //print_r ($artigos);
50    $path = "textos/revdiscursosfo/";
51    $i=0;
52    $id_padrao=seleciona_id_padrao_ref($conecta_vcpc)+1;
53    while ($i< count($artigos)){
54        $arquivo= $artigos[$i]['id'].".txt";
55  //        echo $arquivo.'<br>';
56        //$textoartigo= file_get_contents($path.$arquivo);
57        $texto = fopen ($path.$arquivo, "r", 0);
58        $textoartigo= fread($texto, filesize($path.$arquivo));
59        fclose ($texto);
60        $palavras= separa_palavras ($textoartigo);
61        $sequencia= 1;
62        $x=0;
63        while ($x < count($palavras)){
64            //padrao 'tais como'     //(1) SN (tais como | como) SN { , SN ... , } (e | ou) SN
65            if ((trim ($palavras[$x]) == 'tais') AND (trim ($palavras[$x+1]) == 'como')){
66                $y= $x-20; //antecessor do padrão
67                if ($y < 0) {
68                    $y=0;
69                }
70                $pal_ant_ori='';
71                $pal_ant_pv='';
72                $pal_ant_pl_='';
73                $pal_ant_suf_='';
74                while ($y<$x) {
75                    $pal_utf8= convert_encoding($palavras[$y], 'UTF-8');
76                    $pal_ant_ori.= trim($pal_utf8).' ';
```

Fonte: Elaborado pelo autor (2020).

Como resultado da execução do algoritmo de identificação dos padrões, foram recuperados um total de 7223 excertos, sendo 4690 da revista Informação & Informação, 1552 da revista Informação@Profissões e 981da revista Discursos Fotográficos. No Quadro 23, relacionam-se os resultados por termo-chave utilizado para a identificação dos padrões nos textos de cada artigo em sua respectiva revista.

Quadro 23 – Resultados das quantidades de excertos por padrão

Revista	Termo-chave - Padrão	Quantidades
Discursos Fotográficos	como	850
Discursos Fotográficos	especialmente	7
Discursos Fotográficos	tal(is)	74
Discursos Fotográficos	particularmente	5
Discursos Fotográficos	tais como	5
Discursos Fotográficos	sobretudo	11
Discursos Fotográficos	e outro(s)	6
Discursos Fotográficos	incluindo	5
Discursos Fotográficos	principalmente	15
Discursos Fotográficos	em especial	2
Discursos Fotográficos	ou outro(s)	1
Informação & Informação	como	3824

Continua

Continuação

Informação & Informação	incluindo	42
Informação & Informação	tal(is)	394
Informação & Informação	principalmente	150
Informação & Informação	em especial	30
Informação & Informação	sobretudo	30
Informação & Informação	especialmente	87
Informação & Informação	e outro(s)	49
Informação & Informação	tais como	58
Informação & Informação	em particular	9
Informação & Informação	ou outro(s)	10
Informação & Informação	particularmente	7
Informação@Profissões	como	1307
Informação@Profissões	tais como	7
Informação@Profissões	tal(is)	121
Informação@Profissões	principalmente	54
Informação@Profissões	sobretudo	8
Informação@Profissões	e outro(s)	20
Informação@Profissões	especialmente	15
Informação@Profissões	incluindo	10
Informação@Profissões	em especial	8
Informação@Profissões	ou outro(s)	2

Fonte: Dados da pesquisa.

Soma-se aos resultados desta etapa um grande *corpus* linguístico a partir da busca e recuperação dos excertos retirados dos textos dos artigos científicos demarcados por 20 palavras que antecedem o termo-chave utilizado para localização do possível padrão, e 20 que o sucedem.

6.3 IDENTIFICAR OS PADRÕES DE HEARST A PARTIR DAS PALAVRAS-CHAVE NÃO COMPATIBILIZADAS - ETAPA 3

Os dados resultantes da execução da ação de 1 são decorrentes da execução do algoritmo de busca nos excertos dos termos das palavras-chave não compatibilizadas. O resultado é representado pela codificação do algoritmo de busca dos termos no texto do artigo (que já constam como registros no banco de dados - Figura 29). A partir do recorte dos artigos publicados nos anos de 2018 e 2019 de cada revista, listam-se as palavras-chave não

compatibilizadas internamente no algoritmo (Figura 29) e a partir delas realizam-se as buscas pelos textos dos artigos, por meio dos registros dos excertos. Utiliza-se como operador o *LIKE* em linguagem SQL.

Figura 29 - Codificação do algoritmo de busca de excertos

```
//COMPARATIVO DA PC COM EXCERTO ANTECESSOR AO PADRÃO SEM PALAVRAS VAZIAS
   $strSQL_padrao = "SELECT * FROM `vcpc_int_textoartigos_padrao`
       WHERE (`id_artigo` =".$artigos[$i]['id'].") AND (`excerto_ant_pv` LIKE '%".
           $pcs[$l]['stem_spv']. "%')";
   $rs_padrao = mysql_query($strSQL_padrao, $conecta_vcpc) or die(mysql_error());
   if (mysql_num_rows ($rs_padrao)> 0){
       while ($row = mysql_fetch_array($rs_padrao)){
           $html.= '<tr><td>'.$artigos[$i]['id'].'</td>';
           $html.= '<td>'.$pcs[$l]['id'].'</td>';
           $html.= '<td>'.$pcs[$l]['termo'].'</td>';
           $html.= '<td>'.$row['id_padrao'].'</td>';
           $html.= '<td>'.$row['excerto_ant'].'</td>';
           $html.= '<td>'.$row['padrao'].'</td>';
           $html.= '<td>'.$row['excerto_suc'].'</td>';
           $html.= '<td>ANT_PV</td></tr>';
       }
   }
```

Fonte: Elaborado pelo autor (2020).

O algoritmo representado na Figura 29, executado para a base de dados de cada revista, resultou em uma planilha para as respectivas revistas, num total de 93 excertos da revista Informação@Profissões, 387 da revista Discursos Fotográficos e 662 da revista Informação & Informação. No Quadro 24 apresentam-se os resultados quantitativos.

Quadro 24 – Resultados das análises dos excertos

Revista	Categorização	Quantidades
Discursos Fotográficos	CORRESP_FORTE	1
Discursos Fotográficos	CORRESP_FRACO	14
Discursos Fotográficos	N_CORRESP_FORTE	2
Discursos Fotográficos	N_CORRESP_FRACO	26
Discursos Fotográficos	N_CORRESP_P	344
Informação@Profissões	N_CORRESP_FORTE	5
Informação@Profissões	N_CORRESP_FRACO	3
Informação@Profissões	N_CORRESP_P	85
Informação & Informação	CORRESP_FORTE	2
Informação & Informação	CORRESP_FRACO	26
Informação & Informação	N_CORRESP_FORTE	11
Informação & Informação	N_CORRESP_FRACO	32
Informação & Informação	N_CORRESP_P	591

Fonte: Dados da pesquisa.

Os resultados das análises dos excertos foram categorizados para sinalizar onde foi possível a extração da relação hierárquica, bem como os excertos que não corresponderam aos padrões. Vale ressaltar que na quantificação dos excertos, para as análises apresentadas na sequência, foram desconsideradas as duplicações, visto que um excerto pode constar uma ou várias vezes, levando em conta o vínculo da consulta com a palavra-chave dos artigos e também o termo-chave para identificação dos padrões.

Padrão SN (tais como | como) SN { , SN ... , } (e | ou) SN

O padrão consta com um "ou exclusivo", que somente será verdade onde existir um ou o outro, por exemplo, não poderá existir registros com "tais como como". Portanto foram localizados três registros desse padrão a partir do termo-chave "tais como" na revista Informação & Informação e somente um na Informação@Profissões. Observa-se a precisão desse padrão, apesar de uma quantidade mínima e classificado como <N_CORRESP_FORTE> <N_CORRESP_FRACO> e <CORRESP_FRACO>. O primeiro sintagma nominal dá origem ao hiperônimo e os demais sintagmas nominais, após o termo-chave, indicam os hipônimos.

Excerto ...SOCs, <u>tais como</u> tesauros, taxonomias, sistemas de classificação e listas de cabeçalhos de assunto...

Representação sistemática
SOCs
- ***tesauros***
- ***taxonomias***
- ***sistemas de classificação***
- ***listas de cabeçalhos de assunto***

Excerto ... ambientes virtuais de socialização e compartilhamento de informações on-line, <u>tais como</u> as wikis, mídias sociais, blogs e outros....

Representação sistemática
Ambientes virtuais de socialização
- ***wikis***
- ***mídias sociais***
- ***blogs***
- ***outros***

Nesse contexto, em que se discutem os resultados da identificação dos padrões de Hearst por meio de um termo-chave, vale destacar a diferença de ambos. O termo-chave é parte do padrão, por meio dele as consultas foram estruturadas, em linguagem SQL, e as executadas no banco de dados. O padrão de Hearst é a estruturação dos sintagmas nominais que antecedem e sucedem o termo-chave, representados por meio de fórmulas.

Com o termo-chave "como" obteve-se uma quantidade maior categorizada como <CORRESP_FORTE> e <N_CORRESP_FORTE>, sendo cinco registros de 545 (outras categorizações) na revista Informação & Informação, dois registros na Informação@Profissões, de 77, e dois registros na Discursos Fotográficos, de 352. Observa-se a imprecisão quando utilizado esse termo-chave, a partir do qual se recupera uma quantidade grande de dados, porém uma porcentagem pequena corresponde ao padrão, sendo na maioria dos registros cercada de sintagmas verbais.

Excerto ...recintos públicos <u>como</u> restaurantes, comércio, banheiros, praças públicas.

Representação sistemática
Recintos públicos
restaurantes
comércio
banheiros
praças públicas

O segundo padrão (tal(is) SN como {(SN,)*(ou|e)} SN) obteve várias ocorrências do termo-chave, porém nenhum registro foi possível de correspondência com o padrão e, consequentemente, sem extração de relações hierárquicas expressivas, como exemplo, segue um excerto.

Padrão tal(is) SN como {(SN,)*(ou|e)} SN

Excerto ... tal práxis <u>como</u> parte integrante do que vem se denominando....

No quarto padrão (SN {, SN}* {,} e outro(s) SN), as ocorrências não tiveram nenhuma correspondência com as palavras-chave, porém, para efeitos de exemplificação, há dois

excertos que correspondem ao padrão e, consequentemente, existem relações hierárquicas, porém não suficientes para compor uma sistemática das palavras-chave.

Padrão SN {, SN}* {,} e outro(s) SN
Excerto ... os bibliotecários <u>e outros</u> profissionais da informação....
Excerto ... ato fotográfico <u>e outros</u> ensaios...

O terceiro (SN {, SN}* {,} ou outro(s) SN) e o quinto (SN {,} incluindo {SN,}*{ou|e} SN) padrão não obtiveram registros recuperados a partir das palavras-chave não compatibilizadas dos artigos.

Padrão SN {, SN}* {,} ou outro(s) SN
Padrão SN {,} incluindo {SN,}*{ou|e} SN

O sexto padrão (SN {,} especialmente {SN,}*{ou|e} SN) e suas variações (principalmente, particularmente, em especial, em particular, de maneira especial, sobretudo) obtiveram algumas ocorrências, porém pouco expressivas e sem possibilidade de extração de relações. No primeiro excerto não é possível estabelecer relação entre "bibliotecas tradicionais" e "países econômica e tecnologicamente mais desenvolvidos". Por outro lado, no segundo excerto é possível estabelecer uma relação entre "Web Semântica" e "Linked Data".

Padrão SN {,} especialmente {SN,}*{ou|e} SN
Excerto ... bibliotecas tradicionais <u>especialmente</u> nos países econômica e tecnologicamente mais desenvolvidos....
Excerto ...Web Semântica, <u>em especial</u> do Linked Data...

Diante dos resultados dessa etapa de localização, identificação e análise dos excertos a partir das palavras-chave não compatibilizadas, é possível afirmar que foram expressivas a qualidade e, consequentemente, a possibilidade de extração de relações hierárquicas. Vale destacar, no que se refere a quantidade, que se optou por um recorte do *corpus*. Outra expressividade é a complexidade envolvida na extração dos padrões e análise das ocorrências, a partir de um *corpus* textual puro. Sintetizam os resultados, no entanto, a precisão do primeiro padrão considerando o termo-chave "tais como", considerando a ocorrência e a relação hierárquica extraída, ora o mesmo padrão com o segundo termo-chave "como" a imprecisão é referenciada na quantidade de ocorrências e suas respectivas possibilidades de extração de relações.

6.4 MAPEAR, POR MEIO DOS PADRÕES DE HEARST (1992; 1998), A SISTEMÁTICA DAS PALAVRAS-CHAVE COM O VOCABULÁRIO CONTROLADO E COMPATIBILIZAR COM OS NÍVEIS DE CORRESPONDÊNCIA - ETAPA 4

Os resultados da etapa 3 totalizaram uma quantidade expressiva em relação ao *corpus*, sendo dois registros de padrões na revista Informação & Informação e um registro na revista Discursos Fotográficos categorizados como <CORRESP_FORTE>, que indica uma relação hierárquica forte a partir do padrão de Hearst. Consequentemente, ampliou-se para a análise das demais categorias na ordem <CORRESP_FRACO> e <N_CORRESP_FORTE>, com vistas a mapear mais palavras-chave (Quadro 25 e Quadro 26), mesmo constando uma possível relação hierárquica fraca da palavra-chave por meio do padrão Hearst. Por outro lado, a análise das categorias <N_CORRESP_FORTE> foi realizada para explicar as adjacências das palavras-chave no em torno do padrão (Quadro 25).

Quadro 25 – Mapeamento sistemático de palavras-chave revista Informação & Informação

Palavra-Chave	**Termo TBCI**	**Níveis de correspondência (Neville, 1970)[10]**	**Categorização Padrões**
Revisão sistemática de literatura	Modelos de Análise	Caso 4	CORRESP_FORTE
Análise Categoria	Categorias	Caso 4	CORRESP_FORTE
Investigação Criminal	-	Caso 3	CORRESP_FRACO
Respeito à diversidade	-	Caso 3	CORRESP_FRACO
Narrativa	-	Caso 3	CORRESP_FRACO
Colaboração Científica	comunicação científica produtividade científica	Caso 4	CORRESP_FRACO
Política de indexação	indexação indexação temática	Caso 4	CORRESP_FRACO

Continua

[10] Caso 1: Correspondência exata entre os termos dos tesauros.
Caso 2: Diferentes sinônimos são usados para um mesmo conceito.
Caso 3: O tesauro de origem tem um termo para o conceito e no outro tesauro não existe.
Caso 4: O termo no tesauro de origem existe no outro tesauro, mas com um termo mais geral.
Caso 5: O tesauro de origem usa termos pré-coordenados e o outro tesauro usa termos pós-coordenados.
Caso 6a: O tesauro de origem distingue homônimos e o outro tesauro não.
Caso 6b: O tesauro de origem não distingue homônimos e o outro tesauro faz distinção.
Caso 7: Um tesauro utiliza descritores separados para distinguir um termo usado em sentidos diferentes, enquanto o outro tesauro o não faz.
Caso 8: O tesauro fonte utiliza descritores que não são suficientes para esclarecer o conceito a que se referem.
Caso 9: O tesauro fonte contém descritores sinônimos.
Caso 10: O tesauro fonte faz uso de termos com significado apenas para o uso no local de origem.
Caso 11: Um tesauro utiliza um sistema de codificação arbitrário para alguns conceitos.

Continuação

Patrimônio	preservação de documentos	Caso 4	CORRESP_FRACO
Patrimônio cultural	preservação de documentos	Caso 4	CORRESP_FRACO
Encontrabilidade da Informação	depósito de dados	Caso 4	CORRESP_FRACO
Repositórios de Dados	repositórios digitais	Caso 2	CORRESP_FRACO
Movimentos Sociais	-	-	CORRESP_FRACO
Fintech	tecnologias da informação e comunicação	Caso 4	CORRESP_FRACO
SKOS	sistemas de organização do conhecimento	Caso 4	CORRESP_FRACO
Diplomática	gestão de documentos	Caso 4	CORRESP_FRACO

Fonte: Dados da pesquisa.

Quadro 26 – Mapeamento sistemático de palavras-chave revista Discursos Fotográficos

Palavra-Chave	**Termo TBCI e Comunicação Visual**	**Níveis de correspondência (Neville, 1970)**	**Categorização Padrões**
História	História [TBCI]	Caso 1	CORRESP_FORTE
Retrato	-	Caso 3	CORRESP_FRACO
Imagem	Fotografia e história [CV]	Caso 4	CORRESP_FRACO
Real	Fotografia [CV]	Caso 4	CORRESP_FRACO
Cor	Fotografia [CV]	Caso 4	CORRESP_FRACO
Discurso	Fotojornalismo [CV]	Caso 4	CORRESP_FRACO
Produção de sentidos	-	Caso 3	CORRESP_FRACO

Fonte: Dados da pesquisa.

6.5 DESENVOLVER O MAPEAMENTO REVERSO DAS PALAVRAS-CHAVE COMPATIBILIZADAS PELOS PROCESSOS DE SANTOS (2015) A PARTIR DOS NÍVEIS DE CORRESPONDÊNCIA. - ETAPA 5

A base de dados gerada pela VCPC *Tools* (SANTOS, 2015) contém as tabelas: Artigos; Palavras-chave; VC - Periódico (termos, vínculo com a PC, relacionamentos temporariamente construídos para descrever os relacionamentos dos termos do VC - periódico).

Essa base de dados foi coletada separadamente de cada um dos periódicos científicos eletrônicos da área de Comunicação e Informação da UEL, dados apresentados no Quadro 27, *corpus* de análise e resultados desta etapa. Na coluna <Temos TBCI> são demonstradas as quantidades de termos do vocabulário de referência utilizado para compatibilizar as palavras-chave. Na coluna <Termos Tratados> estão os descritores originados do tratamento das palavras-chave dos artigos, já as colunas <Termos Loc> e <Termos Nome> referem-se à localização e a nomes próprios, respectivamente. Nas três últimas colunas, está a totalidade de palavras-chave beneficiadas pela compatibilização, considerando-se na contagem a atribuição nos respectivos artigos.

Quadro 27 - Termos x Palavras-chave da base da VCPC *Tools*

Periódico	Termos TBCI	Termos Trata-dos	Termos Loc	Termos Nome	Total	PC em Trata-mento	PC Tratadas	PC Total
Informação & Informação (737 artigos)	322	475	9	2	808	389	1015	1404
Informação@ Profissões (140 artigos)	60	-	-	-	60	103	65	168
Discursos Fotográficos (425 artigos)	-	54	-	-	54	892	74	966

Fonte: Dados da pesquisa.

Os resultados do mapeamento reverso são os procedimentos de identificação dos níveis de correspondência das palavras-chave compatibilizadas com um termo do vocabulário de referência. Portanto o entendimento do termo designado como mapeamento reverso é decorrente da sequência dos processos de interoperabilidade: 1º) mapeamento e 2º) compatibilização. No estudo de caso aplicado, o *corpus* de análise está compatibilizado, ou seja, o conceito de mapeamento como atividade inicial integrou-se nos processos de compatibilização de Santos (2015).

Os resultados da modelagem e execução do mapeamento reverso de palavras-chave compatibilizadas por Santos (2015) aplicado como estudo de caso na revista Informação & Informação e posteriormente nas demais revistas da área de Comunicação e Informação (Informação@Profissões e Discursos Fotográficos) da UEL estão apresentados no Quadro 28,

Quadro 29, Quadro 30, respectivamente, com 10 registros como exemplos. As listagens completas e extratificadas podem ser conferidas nos apêndices.

Quadro 28 – Mapeamento de palavras-chave revista Informação & Informação

Palavra-chave	Termo TBCI	Níveis de correspondência (Neville, 1970)	Mapeamento
AACR2	AACR2	Caso 1	IDENTICO
Abordagem Sociocultural da Ciência da Informação	ciência da informação	Caso 4	I_INDICE
acepção	análise linguística	Caso 4	CORRESP
Acervo Bibliográfico	acervos bibliográficos	Caso 1	IDENTsPL
Acesso	acesso	Caso 1	IDENTICO
...	...	....	...
Web Social Semântica	web semântica	Caso 4	I_INDICE
Website	sítios web	Caso 2	CORRESP
Websites	sítios web	Caso 2	CORRESP
World Wide Web	World Wide Web	Caso 1	IDENTICO
XML	XML	Caso 1	CORRESP

Fonte: Dados da pesquisa.

Os níveis de correspondência de Neville (1970) utilizados foram os Casos 1, 2, 4. Os casos de correspondência pelo caso 1 estão ligados diretamente pelos processos de compatibilização <IDENTICO>, <IDENTsPV>, <IDENTsPL> e <IDENTsSUF>.

Quadro 29 – Mapeamento de palavras-chave revista Informação@Profissões

Palavra-chave	Termo TBCI	Níveis de correspondência (Neville, 1970)	Mapeamento
Acessibilidade	acessibilidade	Caso 1	IDENTICO
Administração	administradores	Caso 1	IDENTsSUF
Arquitetura da Informação	arquitetura de informação	Caso 1	IDENTsPV
Arquivista	arquivistas	Caso 1	IDENTsPL
Arquivologia	arquivologia	Caso 1	IDENTICO
...	...	....	...
Segurança da Informação	segurança da informação	Caso 1	IDENTICO
Sistemas de Informações	sistemas de informação	Caso 1	IDENTsPL
Tecnologias da Informação	tecnologias da informação	Caso 1	IDENTICO
Teologia	teologia	Caso 1	IDENTICO
Web Semântica	web semântica	Caso 1	IDENTICO

Fonte: Dados da pesquisa.

Os casos de correspondência pelo caso 2 são referenciados no tratamento intelectual[11] realizado pelos editores das revistas; neste mapeamento são denominados < CORRESP>.

Quadro 30 – Mapeamento de palavras-chave revista Discursos Fotográficos

Palavra-chave	Termo TBCI	Níveis de correspondência (Neville, 1970)	Mapeamento
Acervos Cinematográficos	Acervos cinematográficos	Caso 1	IDENTICO
Acervos Fotográficos	Acervos fotográficos	Caso 1	IDENTICO
Análise do Discurso	Análise do discurso	Caso 1	IDENTICO
Análise do Discurso	Análise do discurso	Caso 4	I_INDICE
Análise Fílmica	Análise fílmica	Caso 1	IDENTICO
...	...	....	...
Teoria Semiótica	Teoria semiótica	Caso 1	IDENTICO
Teorias da Fotografia	Teorias da fotografia	Caso 1	IDENTICO
Vídeo Comunitário	Vídeo	Caso 4	I_INDICE
Vídeo Documentário	Vídeo documentário	Caso 1	IDENTICO
Vídeo Jockey	Vídeo	Caso 4	I_INDICE

Fonte: Dados da pesquisa.

Os casos de correspondência pelo caso 4 é referenciado pela compatibilização por meio do índice contido aplicado por Santos (2015), mapeamento nominado < I_INDICE>.

Quadro 31 – Mapeamento de palavras-chave

Revista	Caso 1	Caso 2	Caso 4	TOTAL
Informação & Informação	293	44	175	512
Informação@Profissões	58	-	7	65
Discursos Fotográficos	52	-	23	75
TOTAL	403	44	205	652

Fonte: Dados da Pesquisa.

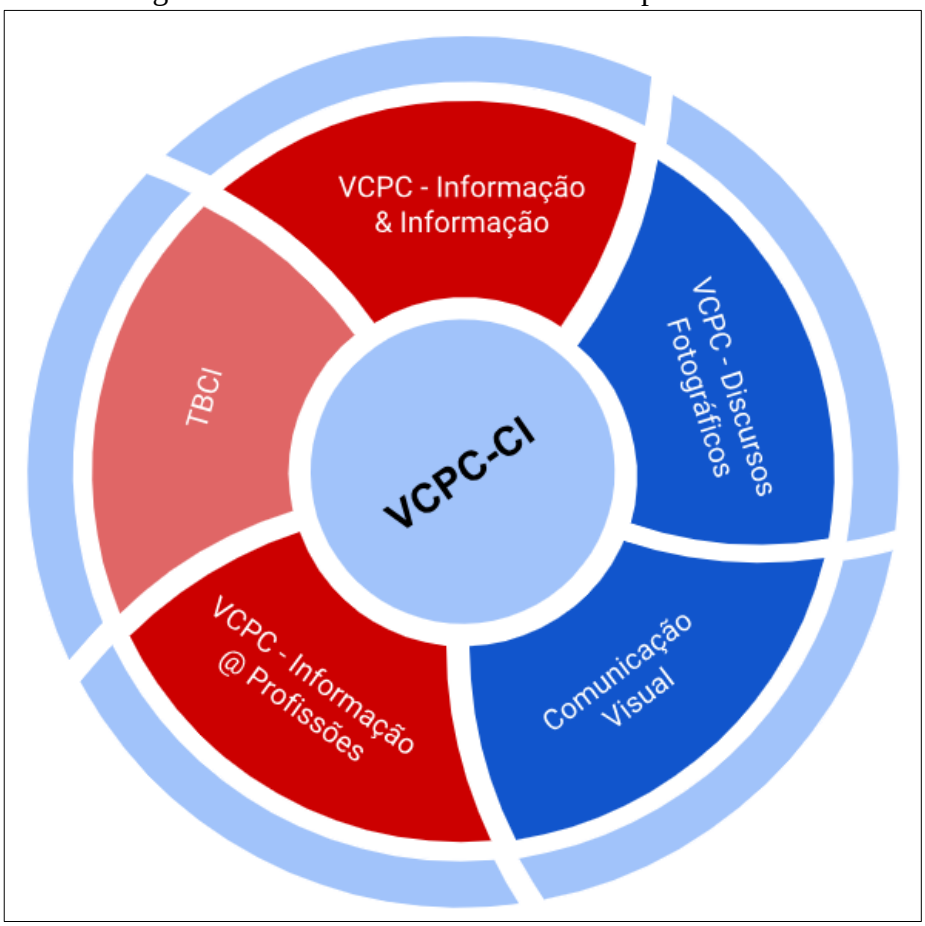
Os resultados da ação de desenvolvimento do modelo conceitual e lógico de interoperabilidade, a partir do *corpus*, analisado, demonstraram o mapeamento dos vocabulários das revistas da área de Comunicação e Informação da UEL, constituídos por meio

[11] O tratamento intelectual das palavras-chave foi realizado pelos bolsistas do Programa de Pós-Graduação em Ciência da Informação da UEL sob a orientação dos editores do periódico. As atividades realizadas constaram no relatório final da bolsista Tânia da Costa Calheiros, detalhado conforme as fases de tratamento, dentre elas as principais são: a) consulta a instrumentos como tesauros, dicionários, glossários da área; b) consulta aos artigos do próprio periódico relacionado com a palavra-chave; c) consulta a outros periódicos da área qualificados pelo Qualis da Capes, no extrato A1; d) opção por manter a palavra-chave atribuída pelo autor, realizando a normalização da palavra-chave como descritor a partir das diretrizes do Manual Técnico para tratamento das palavras-chave. O Manual Técnico até o momento de finalização da tese não se encontra publicado, por esse motivo não consta como referência.

da compatibilização das palavras-chave atribuídas aos artigos científicos. Na Figura 30 apresenta-se o modelo conceitual, por meio da visualização gráfica, modelado a partir dos instrumentos de controle de vocabulário e uma representação da interoperabilidade. O modelo integrativo dos vocabulários controlados dos periódicos científicos eletrônicos da área de Comunicação e Informação denomina-se Vocabulário Controlado dos Periódicos Científicos de Comunicação e Informação (VCPC-CI), e se deu por meio dos vocabulários dos periódicos Informação & Informação, Informação@Profissões e Discursos Fotográficos na ferramenta VCPC *Tools.*

O vocabulário controlado do periódico Discursos Fotográficos é compatibilizado com o vocabulário Comunicação Visual, o qual foi desenvolvido a partir da compilação das palavras-chave do próprio periódico e disponibilizado no TemaTres. No modelo conceitual de interoperabilidade, Figura 30, está representada a interoperabilidade entre os vocabulários controlados nos quesitos sintáticos e semânticos (vocabulários gerenciados por sistema – TemaTres).

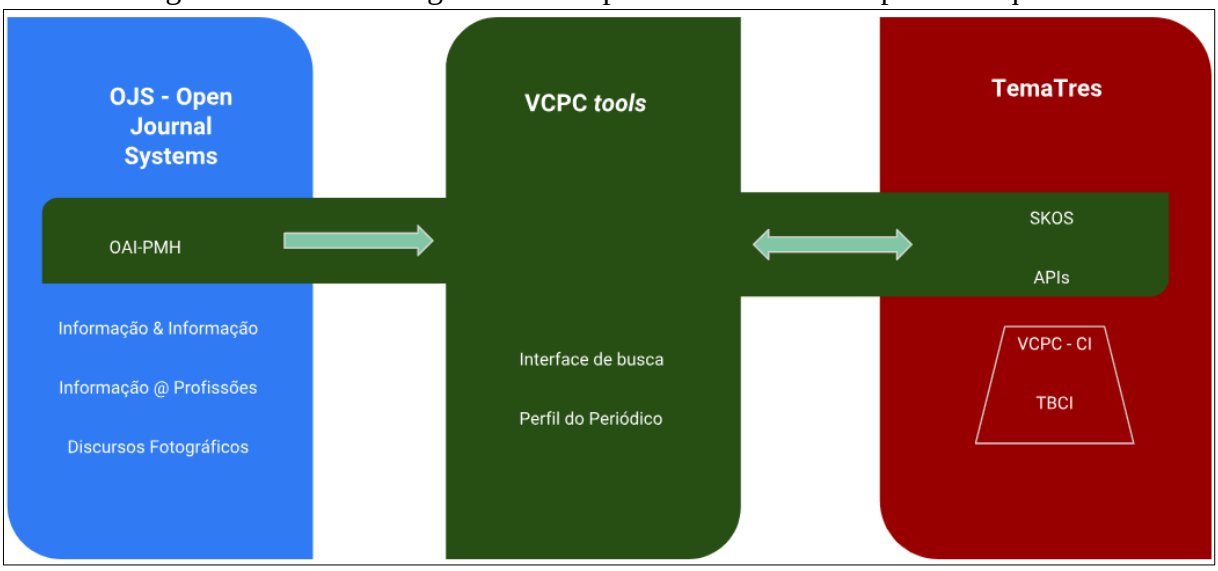
Figura 30 – Modelo conceitual de interoperabilidade



Fonte: Elaborado pelo autor (2020).

O instrumento que consta no centro da Figura 30 é o vocabulário integrativo dos periódicos, denominado VCPC-CI, as cores trabalhadas no modelo conceitual são de aproximação dos vocabulários controlados integrados. Por exemplo, o vermelho e sua variação indicam a compatibilidade exercida entre esses vocabulários. Neste caso, o TBCI foi utilizado como um vocabulário de referência para a constituição do VCPC das revistas Informação@Profissões e Informação & Informação. A implementação do modelo lógico de interoperabilidade perpassa pelos aspectos técnicos das ferramentas de *software* utilizadas. O modelo lógico (Figura 31) de interoperabilidade inclui, além dos aspectos tratados no modelo conceitual (Figura 30) de interoperabilidade sintática e semântica, os requisitos estruturais e sistêmicos para realizar a conexão entre os sistemas VCPC *Tools* e o TemaTres.

Figura 31 – Modelo lógico de interoperabilidade – Protótipo de Pesquisa



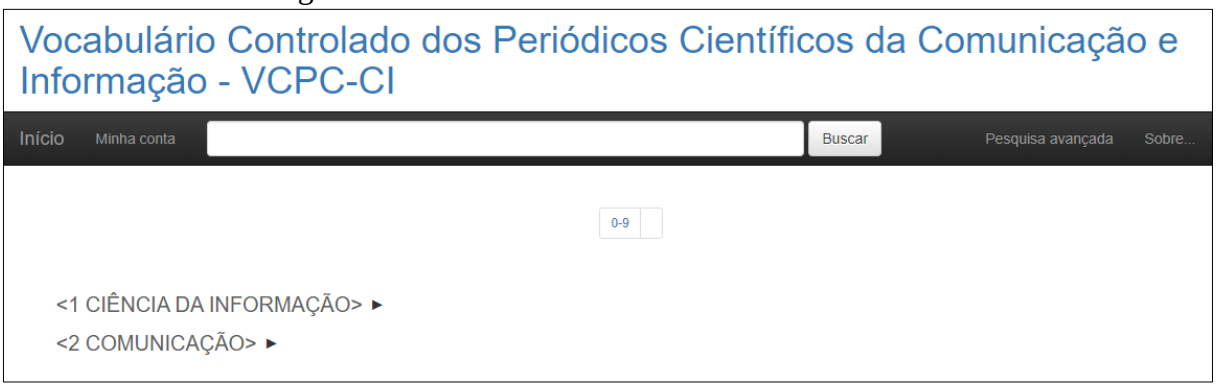
Fonte: Elaborado pelo autor (2020).

O OJS é representativo neste modelo lógico (Figura 31), com a função de fornecer os metadados dos artigos sob sua gerência e hospedados nos repositórios dos respectivos periódicos. Nesse contexto, esta interoperabilidade já é atendida pela VCPC *Tools*. A interface de busca e o perfil do periódico são requisitos funcionais mapeados na ferramenta VCPC *Tools*. O grande desafio é o protótipo de uma interface entre o TemaTres e a VCPC *Tools*. Esse desafio é condicionado à necessidade de que os sistemas tenham API e *Webservice* para realizar o processo de troca de dados, ou seja, a interoperabilidade. No caso do TemaTres, está previsto em seus requisitos funcionais a API, por outro lado, na VCPC *Tools* não foi mapeada esta funcionalidade. Por esse motivo o protótipo da interface apresentado na Figura 37 tornou-se um desafio de planejamento para a realização da interoperabilidade sistêmica com a proposta de um novo requisito funcional na VCPC *Tools*. O TemaTres necessita conter em sua base de

dados o VCPC-CI de modo que para o momento, passa-se a apresentar os resultados do processo de construção do vocabulário no TemaTres.

Tem-se como resultado desse processo a constituição da macroestrutura do VCPC-CI determinada pela subárea de avaliação da Capes Comunicação e Informação, sendo denominada Ciência da informação - CI e Comunicação - CO. A subárea de avaliação da Capes foi utilizada como classificação basilar para o desenvolvimento da macroestrutura (utilizando os dois primeiros níveis). Para tanto, foi importada a macroestrutura no software TemaTres a partir de planilhas elaboradas para organizar seus respectivos termos, designados de metatermos da área de Comunicação (CO) e da área da Ciência da Informação (CI). Na área da CO foi importada a estrutura conceitual elaborada por Santos, Cervantes e Londero (2018). Na área da CI foi importado o plano de classificação do TBCI. Com a macroestrutura do VCPC-CI importada no TemaTres, totalizam-se 65 termos e 61 relações hierárquicas. Estão representados na Figura 32 os metatermos mais gerais da macroestrutura (termos tópicos).

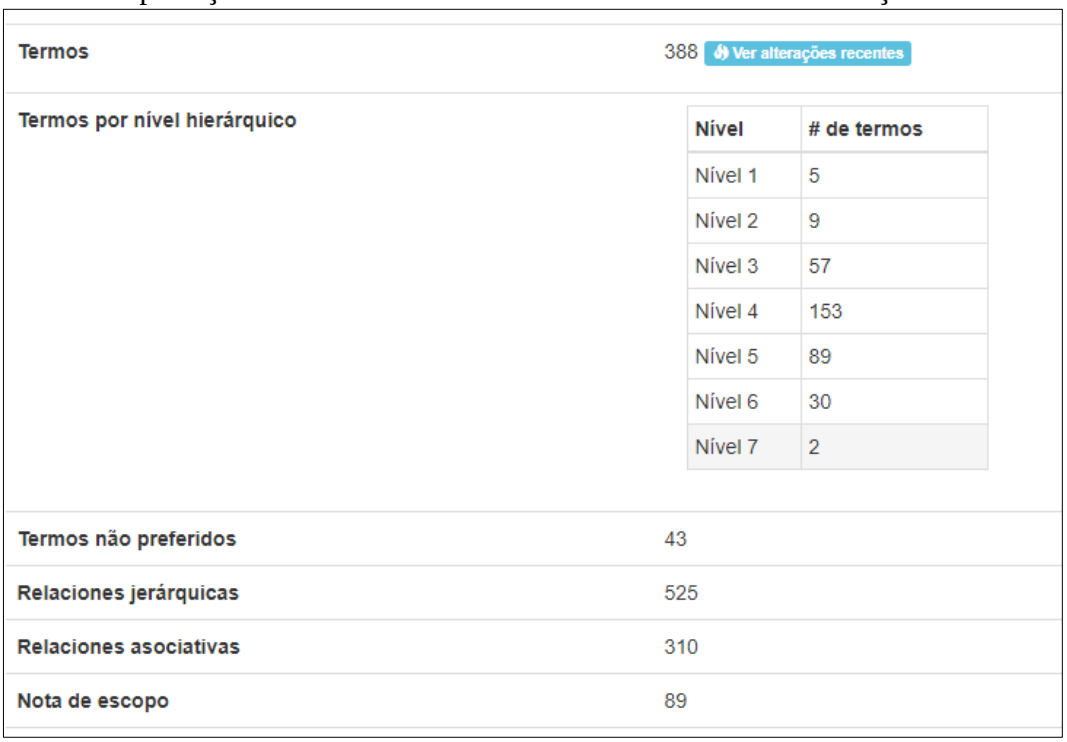
Figura 32 – Interface do VCPC-CI – macroestrutura



Fonte: Elaborado pelo autor (2020).

Após a criação da macro estrutura do VCPC-CI, passa-se a importar os temos que compõem o vocabulário controlado de cada revista. Consideram-se os termos de vocabulários externos em que existem palavras-chave compatibilizadas. Os resultados dessa organização dos termos para importação no vocabulário controlado integrativo dos periódicos são visualizados por meio dos *prints* de parte da interface “sobre” do TemaTres. Na Figura 33 é apresentado o total de termos (n=388), com o destaque para a quantidade de relações hierárquicas (n=525).

Figura 33 - Importação de termos do vocabulário controlado da Informação & Informação



Termos	388 Ver alterações recentes
Termos por nível hierárquico	

Nível	# de termos
Nível 1	5
Nível 2	9
Nível 3	57
Nível 4	153
Nível 5	89
Nível 6	30
Nível 7	2

Termos não preferidos	43
Relaciones jerárquicas	525
Relaciones asociativas	310
Nota de escopo	89

Fonte: Dados da pesquisa.

A importação foi realizada de maneira sequencial, portanto a Figura 34 apresenta a totalização de termos importados, com a execução da inserção dos termos da Informação@Profissões, total de termos (n=403).

Figura 34 - Importação de termos do vocabulário controlado da Informação@Profissões

Termos	403 Ver alterações recentes
Termos por nível hierárquico	

Nível	# de termos
Nível 1	5
Nível 2	9
Nível 3	60
Nível 4	160
Nível 5	93
Nível 6	30
Nível 7	2

Termos não preferidos	44
Relaciones jerárquicas	544
Relaciones asociativas	318
Nota de escopo	108

Fonte: Dados da pesquisa.

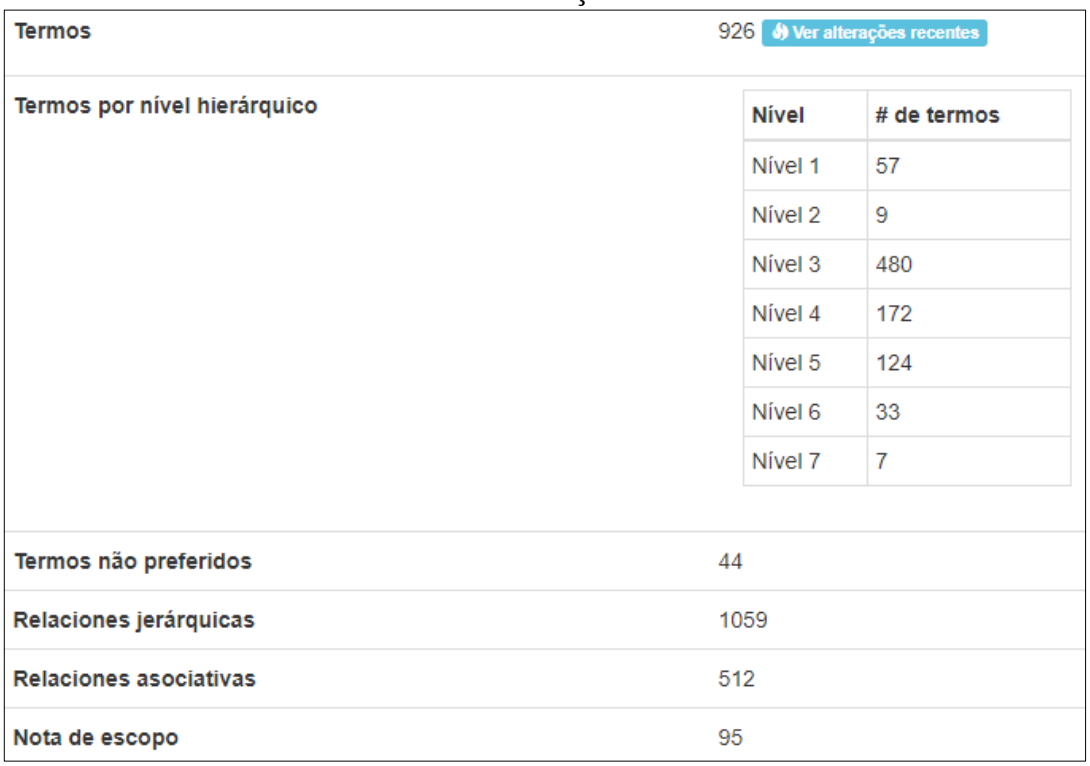
Figura 35 - Importação de termos do vocabulário controlado da Discursos Fotográficos

Termos	457 Ver alterações recentes
Termos por nível hierárquico	**Nível** / **# de termos** Nível 1 / 8 Nível 2 / 15 Nível 3 / 65 Nível 4 / 166 Nível 5 / 119 Nível 6 / 33 Nível 7 / 7
Termos não preferidos	44
Relaciones jerárquicas	637
Relaciones asociativas	512
Nota de escopo	108

Fonte: Dados da pesquisa.

No último processo de importação dos termos derivados de vocabulários externos, como o TBCI, compatibilizados com as palavras-chave, apresenta-se, na Figura 35, o total de termos (n=457). O processo de importação foi finalizado com a importação de 469 palavras-chave tratadas intelectualmente pelos editores da revista e inseridas como descritor no vocabulário da revista Informação & Informação. Essa análise e tratamento intelectual foi realizada somente nessa revista. A previsão de termos no VCPC-CI era de um total de 808 termos, sendo 486 termos tratados e 322 compatibilizados com o vocabulário de referência, o TBCI do periódico Informação & Informação. O mesmo se deu com 60 termos compatibilizados do periódico Informação@Profissões com o TBCI e 54 termos do periódico Discursos Fotográficos. Porém, na importação acontece a verificação de termos já existentes, os quais foram tratados como um termo único, totalizando 926 termos no VCPC-CI, conforme demonstrado na Figura 36.

Figura 36 - Importação de termos tratados intelectualmente da revista Informação & Informação



Termos	926 Ver alterações recentes
Termos por nível hierárquico	

Nível	# de termos
Nível 1	57
Nível 2	9
Nível 3	480
Nível 4	172
Nível 5	124
Nível 6	33
Nível 7	7

Termos não preferidos	44
Relaciones jerárquicas	1059
Relaciones asociativas	512
Nota de escopo	95

Fonte: Dados da pesquisa.

Os registros da interoperabilidade dos termos com o vocabulário de referência, foram realizados de maneira operacional, pois a partir de tentativas de importação no formato SKOS, não foi possível importar os vínculos de interoperabilidade. Neste caso, foram criados dois vocabulários de referência com seus respectivos endereços do *Webservice*, onde se poderá realizar a procura do termo: http://www.uel.br/revistas/informacao/tbci/vocab/services.php [TBCI], http://www.uel.br/revistas/discursosfotograficos/vcpcDF/vocab/services.php [CV]. Por meio desta opção de busca de termos, seleciona-se o termo equivalente nos vocabulários controlados de referência, operação que totalizou 503 termos registrados com seus respectivos vínculos interoperáveis, sendo 451 vínculos com o TBCI e 52 vínculos com o vocabulário Comunicação Visual.

No protótipo da interface de busca integrada (Figura 37), são modelados três principais itens: termos do vocabulário controlado; relacionamentos do termo; e artigos relacionados ao termo selecionado. O funcionamento da interface de busca passa pelos processos de interoperabilidade. Entre a interface e o VCPC-CI no TemaTres foi prevista a troca de informações por meio do *Webservice* e do API do TemaTres. Por outro lado, a interoperabilidade entre a interface de busca e o OJS é realizada por meio de uma API da VCPC *Tools*, sugerida para os seus desenvolvedores.

As requisições das API do TemaTres utilizadas para listar os termos e seus relacionamentos são: <fetchTopTerms>, <fetchDown>, <fetchAlt>, <fetchUp>, <fetchRelated> e <fetchNotes>. O mecanismo de troca de informações acontece por meio de uma requisição e retorno com o resultado. Por exemplo, na requisição <fetchTopTerms> listam-se os termos tópicos do vocabulário controlado a partir da sintaxe de requisição <?task=fetchTopTerms>; o retorno é um arquivo XML com os principais dados: termo e o identificador do termo, dentre outros. No Quadro 32 apresentam-se estas requisições com suas respectivas descrições e resultados.

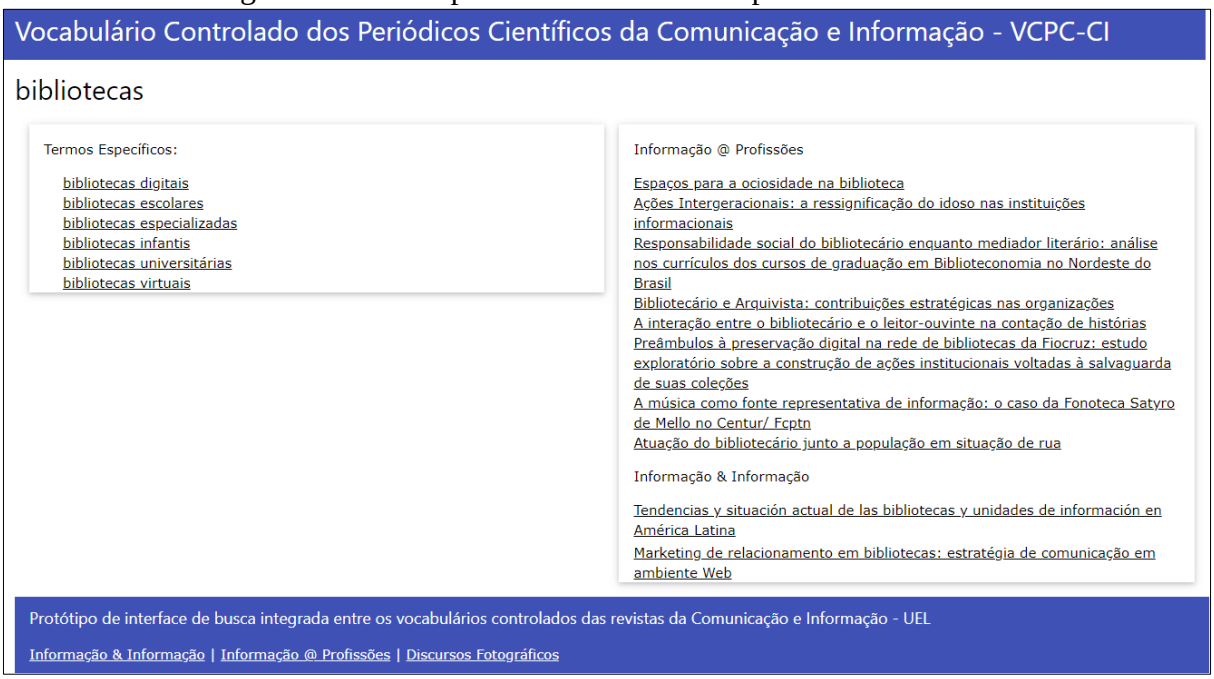
Quadro 32 - Requisições da API do TemaTres utilizadas no protótipo

N	Requisição	Sintaxe	Descrição	Principais dados retorno
1	fetchTopTerms	?task=fetchTopTerms	Lista os termos tópicos.	Termo Identificador do termo
2	fetchDown	?task=fetchDown&arg=1	Lista os termos específicos.	Termo Identificador do termo Tipo relacionamento
3	fetchAlt	?task=fetchAlt&arg=1	Lista os termos não preferidos.	
4	fetchUp	?task=fetchUp&arg=1	Lista os termos mais gerais da estrutura hierárquica.	
5	fetchRelated	fetchRelated&arg=1	Lista os termos relacionados.	
6	fetchNotes	?task=fetchNotes&arg=1	Lista as notas.	Identificador da nota Tipo de nota Nota

Fonte: Dados da pesquisa.

Na Figura 37, apresenta-se o protótipo da interface de busca. O layout foi modelado utilizando-se do estilo disponibilizado pela W3C, o <w3.css>.

Figura 37 - Protótipo interface de busca a partir do VCPC-CI



Fonte: Elaborado pelo autor (2020).

O protótipo da interface de busca procurou-se apresentar com aspectos de facilidade na navegação e na compreensão dos usuários, tanto nos termos do vocabulário controlado, quanto nos artigos aos quais os termos estão vinculados por meio do mapeamento e compatibilização das palavras-chave. Vale destacar o caráter de protótipo: a partir do refinamento e desenvolvimento da versão final para o usuário, poderão ser tomadas decisões sobre o layout, performance, dentre outras características. Por outro lado, sugere-se mecanismos de avaliação a partir do estudo de usuários.

Todos os algoritmos desenvolvidos neste estudo, como auxiliar aos procedimentos operacionais, estão disponíveis em modo texto, o que não impede que, em futuros trabalhos, sejam desenvolvidas as interfaces gráficas para manipulação de dados.

Como proposta para gerenciar a estrutura desenvolvida de interoperabilidade, incluindo-se os aparatos tecnológicos, sugere-se a modelagem e implementação de um aperfeiçoamento da VCPC *Tools*. Nessa proposta, para a manutenção da interoperabilidade VCPC *Tools* com VCPC-CI, a cada lançamento de um volume, mapeiam-se os requisitos funcionais enumerados com RF:

RF01: Coletar os metadados (funcionalidade já existente);

RF02: Compatibilizar as palavras-chave com VCPC-CI (proposta de compatibilidade utilizando os mecanismos de interoperabilidade - API / SKOS);

RF03: Tratar as palavras-chave semiautomáticas (inclusão da instanciação de vínculos de interoperabilidade - verificar a existência do termo no VCPC-CI; adicionar a URI do termo compatível do VCPC-CI; incluir o vínculo na palavra-chave).

RF04: Webservice – API – Disponibilizar os mapeamentos e compatibilizações das palavras-chave e os relacionamentos com os artigos, a fim de operacionalizar a interoperabilidade entre o protótipo da interface de busca.

7 CONSIDERAÇÕES FINAIS

O surgimento da *Web* e o veloz desenvolvimento das tecnologias da informação tornaram sem limites a disponibilização de periódicos científicos por via eletrônica e, como decorrência, altamente complexas as condições de seu acesso. A certeza de dispor de algo acessível, ao alcance das mãos e a qualquer momento, acaba por transformar euforia em frustração, quando os processos de busca trazem resultados aquém das expectativas. Trata-se de um fenômeno que se acentua no cenário atual, em que há uma oferta de textos em vastidão oceânica, ao lado de uma capacidade limitada de conseguir reunir aqueles que realmente podem ser bem utilizados para um dado propósito. O crescimento de informação científica, em especial os artigos científicos eletrônicos, é evidente. São, no entanto, imensas as possibilidades de acesso e igualmente imensas as condições de recolha, o que não significa conseguir captar o que efetivamente é necessário.

A tese defendida neste estudo é de que se necessita de controle de vocabulário interoperável no contexto dos periódicos científicos eletrônicos. A implantação de vocabulários controlados interoperáveis somente terá êxito assegurado se estiver assentada no desempenho de uma ferramenta – evidentemente sujeita a constante aferição e ajuste, por isso não definitiva. O direcionamento deste estudo foi fundado na intenção de construir e apresentar uma proposta teórico-metodológica (um modelo) de interoperabilidade de SOC entre vocabulários controlados de periódicos científicos eletrônicos. A proposta é de natureza teórico-metodológica e poderá ser utilizada para complementar a compatibilização de palavras-chave de artigos científicos eletrônicos, e prosseguir com a manutenção dos vocabulários controlados aplicados em periódicos científicos eletrônicos, em especial aqueles gerenciados pelo OJS.

A questão que deu origem ao estudo desta tese foi expressa pela seguinte pergunta: Como recuperar a informação em periódicos científicos eletrônicos por meio de vocabulários controlados mapeados a partir das palavras-chave em um modelo de interoperabilidade? A hipótese decorrente foi formulada em termos de que os vocabulários controlados, aplicados a periódicos científicos eletrônicos gerenciados pelo OJS, podem ser considerados como um instrumento interoperável para melhorar os processos de busca e recuperação dos artigos, e tornar-se recurso para consultas bem-sucedidas.

Definiu-se, como objetivo geral, apresentar uma proposta teórico-metodológica destinada a subsidiar os processos de implantação e de gestão da interoperabilidade semântica entre SOC e vocabulários controlados aplicados em periódicos científicos eletrônicos. O estudo resultou num protótipo da interface de busca que realiza a integração entre o VCPC-CI

gerenciado pelo TemaTres e o OJS, utilizando como suporte o VCPC *Tools*. A partir da proposta teórico-metodológica desenvolvida e aplicada, a definição de viabilidade de construção de vocabulários controlados interoperáveis apresentou-se como um recurso eficaz para uso como parte integrante de periódicos científicos eletrônicos. Os objetivos específicos foram alcançados conforme descrito a seguir.

O primeiro objetivo específico propunha delimitar os aspectos relacionados aos vocabulários controlados como sistema de organização do conhecimento aplicado aos periódicos científicos eletrônicos. Vale ressaltar que o vocabulário controlado com a finalidade de organização e representação da informação temática para a recuperação da informação em periódicos científicos eletrônicos ainda é escasso na literatura. Há dois trabalhos seminais relacionados com a representação temática de artigos a partir das palavras-chave, dos quais apenas um contém o desenvolvimento de vocabulário controlado em periódicos científicos eletrônicos, recuperados na BDTD. Diante desses escassos resultados, e como ratificação, buscou-se a base *Lattes* de currículos, pela qual seis trabalhos foram recuperados. Três revistas fazem o uso do vocabulário controlado: Informação & Informação, Informação@Profissões e Discursos Fotográficos. Ressalta-se a relevância de ambas as bases de dados no ambiente de pesquisa nacional.

O segundo objetivo específico era explorar a utilização de padrões léxico-sintáticos como subsídios para mapeamento e compatibilização de palavras-chave com a sistemática do vocabulário controlado. A introdução dos aspectos da linguística de *corpus* como ferramenta de auxílio na construção de vocabulários controlados, no que tange ao SOC, deu sustentação no desenvolvimento da proposta teórico-metodológica. A integração desses fundamentos possibilitou uma análise a partir das relações hierárquicas extraídas, para efeitos de incluí-las na sistemática do vocabulário controlado do periódico, e consequentemente realizar o mapeamento e compatibilização das palavras-chave, por meio da sistemática localizada.

Reporta-se, nesse contexto, a aplicação dos padrões linguísticos como suporte e auxílio na tomada de decisão de importar os resultados para um SOC, uma expertise da área de CI integrada com a TI e a Linguística de *Corpus*. Cabe destaque a trabalhos relacionados, em grande parte, com as áreas da computação e linguística, que emergiu pelo menos um trabalho com a área da CI. Evidencia-se uma nova experiência, cujos os resultados obtidos conseguem ensejar novos experimentos e ampliações para contribuir com a proposta em questão, e também com a disseminação de métodos para auxílio na construção de vocabulários controlados, com destaque aos periódicos científicos eletrônicos.

Considera-se relevante destacar, ainda, no contexto desta tese, o protagonismo da aplicação do método de mapeamento por meio dos padrões de Hearst e prospecção do modelo de interoperabilidade. Pode-se afirmar que a aplicação dos padrões de Hearst é uma ferramenta para o auxílio na construção e compatibilização de vocabulários controlados dos periódicos científicos eletrônicos. Além disso, dá suporte para análises de construções textuais e possibilidades de apropriar mecanismos totalmente automatizados para extração das relações hierárquicas. Em grande parte, as palavras-chave já veiculam expressões mais atuais, pelo fato de terem sido originadas por autores recém-ingressados no contexto contemporâneo de sua área. Resta definir o modo como serão conduzidas como termos tratados e integrados ao vocabulário controlado do periódico científico.

No terceiro objetivo específico, a ação era de delimitar o conceito de interoperabilidade no âmbito de sua aplicação em vocabulários controlados para periódicos científicos e seus aspectos relacionados ao modelo de dados e à linguagem, com foco nos processos de mapeamento e compatibilização. Uma limitação decorrente da lacuna nas abordagens do quesito de interoperabilidade entre SOC, encontrada em estudos de casos, reforça a necessidade de subsidiar o desenvolvimento dos processos metodológicos da proposta em questão. Nacionalmente, são quase inexistentes instrumentos interoperáveis, ao passo que, em âmbito internacional, já existem algumas iniciativas. As pesquisas seminais localizadas desenvolvem métodos aplicados com grande frequência às ontologias, o que se justifica pelo fato de o interesse de pesquisa em ontologias ser também afeto à Ciência da Computação, além da CI.

Por meio da *Basel Registerof Thesauri, Ontologies & Classifications* (BARTOC), que é uma base de dados de SOC desenvolvida por *Basel University Library -Switzerland*, é possível um demonstrativo da quantidade de SOC nacionais e internacionais. São 28 registros de SOC nacionais, sendo 7 registros em SKOS (formato interoperável) entre outros, de um total de 2.988 vocabulários indexados na BARTOC. Os registros nacionais representam menos de 1% dos registros internacionais.

O quarto objetivo específico propunha apresentar um modelo de interoperabilidade que inter-relacione semanticamente os vocabulários dos periódicos científicos eletrônicos e o TBCI. A partir desse objetivo, cinco etapas foram delineadas: 1) Reconhecer o ambiente de vocabulário controlado de periódicos científicos eletrônicos gerenciados pelo OJS; 2) Sistematizar o conteúdo dos artigos do periódico científico eletrônico em formato de busca em texto completo a partir da base de dados da VCPC Tools; 3) Identificar os padrões de Hearst a partir das palavras-chave não compatibilizadas; 4) Mapear, por meio dos padrões de Hearst

(1992; 1998), a sistemática das palavras-chave com o vocabulário controlado e compatibilizar com os níveis de correspondência; 5) Desenvolver o mapeamento reverso das palavras-chave compatibilizadas pelos processos de Santos (2015) a partir dos níveis de correspondência.

A execução das ações necessárias para a estruturação da proposta teórico-metodológica em que se realizaram, primeiramente, investigações sobre a temática na BDTD e base *Lattes* de currículos foi fundamental para seleção do *corpus,* objeto de estudo. A análise da disponibilização dos vocabulários controlados dos periódicos científicos eletrônicos possibilitou reconhecer o funcionamento dos vocabulários e suas dependências com a organização da informação para a recuperação dos artigos científicos eletrônicos. Esta pesquisa e coleta inicial exigiram um planejamento das ações como modelagem do projeto lógico do banco de dados, realização de testes de requisição e retorno ao *Webservice* do TemaTres por meio da API.

Integra-se, ao plano de pesquisa, por meio deste objetivo, a etapa de planejamento do modelo conceitual de interoperabilidade nas características sistêmica, estrutural, sintática e semântica. A proposição do modelo de interoperabilidade está contida nos elementos oferecidos para o mapeamento dos vocabulários controlados nos periódicos científicos eletrônicos de maneira integrada: denominou-se VCPC-CI. O mapeamento foi conduzido com o desafio de aplicar os principais mecanismos de compatibilidade de SOC, de Dahlberg e Neville. No entanto os processos de conversibilidade de Neville foram os que se mostraram mais adequados ao desenvolvimento da proposta, considerando-se o foco na compatibilidade entre vocabulários controlados.

O último objetivo específico era aplicar o modelo de interoperabilidade entre vocabulários a um conjunto específico de periódicos (estudo de caso). A execução da proposta teórico-metodológica da pesquisa contou com a aplicação das ações e atividades planejadas, passou por coleta, tratamento e armazenamento em banco de dados relacional, com a finalidade de localizar e identificar os padrões de Hearst. Nos casos de consolidação dos padrões de Hearst, estes passaram a compor, de maneira complementar, os procedimentos de compatibilização apresentados em Santos (2015). As etapas de compatibilização respondem à necessidade do periódico científico eletrônico de poder contar com um vocabulário controlado para organização e representação temática dos artigos científicos eletrônicos.

Salienta-se que os resultados da localização dos excertos relacionados com as palavras-chave tiveram o recorte dos artigos científicos publicados nos anos de 2018 e 2019. A complexidade de desenvolvimento do modelo conceitual de interoperabilidade exige preceitos teórico-práticos de pesquisadores da área de vocabulários controlados, linguística para integrar

a proposta de compatibilização sistemática, modelo de dados para representar o SOC, e demais interfaces de buscas integradas ao sistema gerenciador de vocabulário controlado. As dificuldades encontradas nesta integração entre preceitos teóricos é tema para decisões de pesquisa permanente.

A recuperação da informação a partir de interfaces interoperáveis com vocabulários controlados evidenciou os principais mecanismos de troca de informações entre os sistemas, sendo protocolo de interoperabilidade no OJS, o OAI-PMH; no TemaTres um *Webservise* que responde às requisições por meio de uma API, em suporte XML. Essas questões técnicas e estruturais da interoperabilidade entre os sistemas dão o suporte para o mapeamento dos requisitos funcionais de uma interface de busca, tendo como base a VCPC *Tools*, que dentre as funcionalidades, está a interação entre o OJS e o vocabulário do periódico.

As principais dificuldades encontradas na realização da pesquisa foram a relativa falta de estudos recentes sobre a temática para o desenvolvimento da proposta, considerando-se a dificuldade de localização por meios eletrônicos de materiais antigos e a ausência de elementos técnicos para dar suporte no desenvolvimento de soluções automatizadas. No que tange a esses aspectos técnicos, uma visível dificuldade relacionou-se com a extração ou conversão de textos de arquivos em formato PDF. Essa dificuldade deveu-se às diversificações de tipologias de arquivos nesse formato, considerando-se seus aspecto de portabilidade e codificação, além de características como idioma, tratamento de caracteres especiais e acentuações. Essa ação não foi possível de ser automatizada na pesquisa, tendo em vista que os resultados das conversões esbarraram-se na ininteligibilidade do conteúdo dos artigos que foram *corpus* de análise.

Por outro lado, o registro da interoperabilidade entre os termos do vocabulário controlado dos periódicos com o vocabulário de referência esbarrou-se na impossibilidade de importar no TemaTres esses vínculos. As tentativas foram realizadas a partir de importação em formato SKOS, ou seja, com a exportação dos termos vinculados em SKOS. O próprio gestor de vocabulários controlados não obtive êxito nos testes de importação, considerando a versão 3.0 do *software* em questão. Tornou-se necessária a inclusão dessa interoperabilidade de maneira manual por meio da interface de busca e efetivação de vínculo.

Os encaminhamentos para trabalhos futuros vão ao encontro da necessidade de avaliação do modelo proposto por meio de testes de usabilidade e acessibilidade com a comunidade profissional e usuária da interface de busca da informação. Neste quesito de avaliação, vale acrescentar os mecanismos de busca e recuperação da informação, os quais caracterizam o acesso da informação proposto pelo ODS da agenda de 2030 das Nações Unidas. A análise desse formato de busca, por meio de vocabulários controlados, poderia estimular o

desenvolvimento de estudos com o foco na conversão da necessidade de informação do usuário em estratégias de busca de informação e os resultados, considerando-se os aspectos comparativos.

Conclui-se que a organização e representação temática em artigos científicos publicados em periódicos científicos eletrônicos, disponíveis por meio de seus respectivos repositórios, gerenciados pelo OJS, adquirem uma característica singular quando se considera o contexto do país, sua língua, sua cultura, as condições de geração de conhecimento. Realizar estudos a respeito de vocabulário controlado pode tornar-se um empenho enriquecedor e ao mesmo tempo uma tarefa gratificante, embora árdua, sempre predispondo quem com ela se envolve a situar-se em meio a um universo de tomadas de decisão e auferir para si próprio, e ajudar outros a auferirem, preciosas descobertas. São achados que também podem ajudar no contexto da internacionalização tão estimulada nos tempos atuais. Um vocabulário controlado interoperável com outros vocabulários e interfaces de buscas é fundamental para a eficiência e eficácia da recuperação da informação. Não se realiza qualquer tarefa sem instrumentos. Instrumentos são provisórios e sua atualização tem que ser permanente, por isso o VCPC-CI fica consignado com campo de estudo para futuras pesquisas.

REFERÊNCIAS

ADEBESIN, F.; FOSTER, R.; KOTZE, P.; VAN GREUNEN, D. A review of interoperability standards in e-Health and imperatives for their adoption in Africa. **South African Computer Journal**, Grahamstown, v. 50, n. 1, p. 55-72, 2013.

AGUIAR, F. L.; TALÁMO, M. F. G. M. O Controle de Vocabulário da Linguagem Orgânico-Funcional: Concepção e princípios teórico-metodológicos. **Acervo**, Rio de Janeiro, v. 25, n. 1 jan-Jun, p. 117-138, 2012.

ALVARES, L (Org.). **Organização da informação e do conhecimento**: conceitos, subsídios interdisciplinares e aplicações. São Paulo: B4 Ed., 2012. 248p.

AMERICAN NATIONAL STANDARS INSTITUTE - ANSI; NATIONAL INFORMATION STANDARDS ORGANIZATION - NISO. **Z39:19-2005: guidelines for the construction, format, and management of monolingual controlled vocabularies**. Bethesda: NISO Press, 2005.

ANDRADE, J. **Interoperabilidade e mapeamentos entre sistemas de organização do conhecimento na busca e recuperação de informações em Saúde: estudo de caso em Ortopedia e Traumatologia**. 327 f. 2015. Tese (Doutorado em Ciência da Informação) – Escola de Comunicações e Artes, Universidade de São Paulo (USP), São Paulo, 2015.

ANDRADE, J. LARA, M. L. G. Interoperability and Mapping Between Knowledge Organization Systems**. Knowledge Organization**, Frankfurt, v. 43, n. 2, p. 107-112, 2016.

ANDRADE, M. C.; CERVANTES, B. M. N. A contribuição da organização do conhecimento para a interoperabilidade semântica: alternativas para repositórios institucionais. **Informação@Profissões**, v. 1, n. 1/2, p. 151-169, 2012.

BANDIM, M. A. S. **Indexação automática por atribuição de artigos científicos da área de ciência da informação**. 2017. 139 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Pernanbuco, Recife, 2017.

BARITÉ ROQUETA, M. Sistemas de organización do conhecimento: uma tipologia atualizada. **Informação & Informação**, Londrina, v. 16, n. 2, p. 122-139, dez. 2011. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/9952. Acesso em: 6 jun. 2015.

BARITÉ, M. **Diccionario de Organización del Conocimiento**: Clasificación, Indización Terminología. 6ª ed. Montevideo: CSIC, 2015.

BASÉGIO, T. L. **Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil**. 2007. 124 f. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2007.

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

BODÊ, E. Documento digital e preservação digital: algumas considerações conceituais. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 9, n. 2, p. 503-516, 2016.

BREWSTER, C.; WILKS, Y. Ontologies, Taxonomies, Thesauri: Learning from Texts. In: The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop, 2004. **Proceedings** [...] London: Kings College,2004.

BREWSTER, C.; CIRAVEGNA, F.; WILKS, Y. User-centred ontology learning for knowledge management. In: INTERNATIONAL CONFERENCE ON APPLICATION OF NATURAL LANGUAGE TO INFORMATION SYSTEMS, 2002, Estocolmo. **Anais** [...] Estocolmo: Springer, Berlin, Heidelberg, 2002. p. 203-207.

BRUIJN, J.; MARTIN-RECUERDA, F.; MANOV, D.; EHRIG, M**. State-of-the-art survey on Ontology Merging and Aligning**. Innsbruck: Technical report, SEKT project D4.2.1, 2004.

BURKE, P. **Uma história social do conhecimento**: de Gutenberg a Diderot. Rio de Janeiro: Jorge Zahar, 2003.

CÂMARA TÉCNICA DE DOCUMENTOS ELETRÔNICOS**. Glossário**: Documentos. 6ª ed. Rio de Janeiro: Conselho Nacional de Arquivo – CONARQ, 2014. Arquivísticos Digitais. Disponível em: http://conarq.arquivonacional.gov.br/images/ctde/Glossario/2014ctdeglossario_v6_public.pdf. Acesso em: 28 jan. 2020.

CAMPOS, L. F. B. Metadados digitais: revisão bibliográfica da evolução e tendências por meio de categorias funcionais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 12, n. 23, p. 16-45, 2007.

CAMPOS, M. L. A. A problemática da compatibilização terminológica e a integração de ontologias: o papel das definições conceituais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 6, 2005, Florianópolis. **Anais** [...] Florianópolis: ANCIB; UFSC, 2005.

CAMPOS, M. L. A. Integração de ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 7, 2006, Marília. **Anais** [...] Marília: ANCIB; UNESP, 2006.

CAMPOS, M. L. A.; CAMPOS, M. L. M.; *DÁVILA, A. M. R.; GOMES, H. E.; CAMPOS, L. M.; LIRA, L.* Aspectos metodológicos no reuso de ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 3, n. 1, mar. 2009. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/820. Acesso em: 28 jul. 2019.

CARVALHO, E. M. F. de. **Metodologia de construção de um glossário bilíngue com base em um *corpus* de domínio técnico**. 2007. 81 f. Dissertação (Mestrado) - Curso de Estudos da Tradução, Estudos da Tradução, Universidade Federal de Santa Catarina, Florianópolis, 2007.

CARVALHO, S. A. L. **Terminologia e documentação**: um estudo terminográfico sobre performance musical. 2013. 188 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

CATARINO, M. E.; CERVANTES, B. M. N.; ANDRADE, I. A. A representação temática no contexto da web semântica. **Informação e Sociedade**, João Pessoa, v. 25, n. 3, p. 105-116, 2015.

CERVANTES, B. M. N. **Contribuição para a terminologia do processo de inteligência competitiva:** estudo teórico e metodologico. 2004. 183 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2004.

CIMIANO, P.; MADCHE, A.; STAAB, S.; VOLKER, J. Ontology Learning. In. STAAB, S.; STUDER, R. (Eds.). **Handbook on Ontologies**. 2 ed. New York: Springer Dordrecht Heidelberg London, 2019.

CLARKE, S. G. D.; ZENG, M. L. From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modeling. **Information Standards Quarterly**, Baltimore, v. 24, n. 1, p. 20-26, Winter 2012.

COPPENS, S.; DEBEVERE, P.; MANNENS, E. **Unified Thesaurus feasibility study**: Possibilities, Representation and Design. Pittsburgh: MMLab, ITEC, IBCN. 2014.

CÔRTES, P. L. Considerações sobre a evolução da ciência e da comunicação científica. ln: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. (Org.). **Comunicação e produção científica**: contexto, indicadores e avaliação. São Paulo: Algellara, 2006. p. 33-56.

CUNHA, M. B.; CAVALCANTI, C. R. O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, 2008.

CURRÁS, E. **Ontologias, taxonomias e tesauros em teoria de sistemas e sistemática**. Tradução de Jaime Robredo. Brasília: Thesaururs, 2010.

DAHLBERG, I. Conceptual compatibility of ordering systems**. Knowledge Organization**, Frankfurt, v. 10, n. 2, p. 5-8, 1983.

DAHLBERG, I. Towards estabilishment of compatibility between indexing languages. **Knowledge Organization**, Frankfurt, v. 8, n. 2, p. 88-91, 1981.

DALTIO, J. **Aondê: um serviço Web de ontologias para interoperabilidade em sistemas de biodiversidade**. 107 f. 2007. Dissertação (Mestrado em Ciência da Computação) – Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2007.

DIAS, C. A. Hipertexto: evolução histórica e efeitos sociais. **Ci. Inf.**, Brasília, v. 28, n. 3, p. 269-277,1999. Disponível em: http://dx.doi.org/10.1590/S0100-19651999000300004. Acesso em: 11 Jul. 2019.

DIAS, G. D. **A organização temática da informação em periódicos científicos eletrônicos**: atribuição de palavras-chave na biblioteconomia e ciência da informação. 2012. 159 f. Dissertação (Mestrado em Gestão da Informação) - Universidade Estadual de Londrina, Londrina, 2012.

DISCURSOS FOTOGRÁFICOS, 2019. Disponível em: http://www.uel.br/revistas/uel/index.php/discursosfotograficos. Acesso em: 11 ago. 2019.

DZIEKANIAK, G. V. A organização da informação e a comunicação científica: implicações para os profissionais e usuários da informação. **Em Questão**, Porto Alegre, v. 16, n. 1, p. 45-59, 2010. Disponível em:http://basessibi.c3sl.ufpr.br/brapci/v/a/8958. Acesso em: 01 ago. 2017.

FELICÍSSIMO, C. H. **Interoperabilidade semântica na Web:** uma estratégiapara o alinhamento taxonômico de ontologias. 180 f. 2004. Dissertação (Mestrado em Informática) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2004.

FERNEDA, E. **Recuperação da Informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 147 f. Tese (Doutorado em Ciência da Comunicação) – Universidade de São Paulo, São Paulo, 2003.

FERNEDA, E.; DIAS, G. A. A Lógica Fuzzy Aplicada à Recuperação de Informação. **InterScientia**, João Pessoa, v. 1, n. 1, p. 51-65, 2013. Disponível em:http://hdl.handle.net/11449/114971. Acesso em: 03 ago. 2019.

FREITAS, M. C. **Elaboração automática de ontologias de domínio**: discussão e resultados. 142 f. 2007. Tese (Doutorado em Letras) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

FREITAS, M. C.; QUENTAL, V. S. D. B. Subsídios para a elaboração automática de taxonomias. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 27, Workshop em Tecnologia da Informação e da Linguagem Humana, 5, 2007, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: SBC, 2007.

FUJITA, M. S. L.; SANTOS, L. B. P. DOS; CRUZ, M. C. A.; MOREIRA, W. Avaliação das características do Tema 3 e Multites para o controle de autoridades nas bibliotecas universitárias. **Scire**, Zaragoza, v.23, n.2, p.71-81, jul.-dic. 2017.

LEIVA, I. G. **La automatización de la indización de documentos**. Murcia: Ediciones Trea, 1999.

GALVINO, C. C. T. **A arte de indexar artigos de periódicos**: a política de indexação da seção de periódicos da Biblioteca Central da UFPB. 2012. 88 f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal da Paraíba, João Pessoa, 2016.

GILLAM, L.; TARIQ, M.; AHMAD, K. Terminology and the construction of ontology. **Terminology**, Amsterdam, v. 11, n. 1, p. 55-81, 2005.

GONÇALVES, A.; RAMOS, L. M. S. V. C.; CASTRO, R. C. F. Revistas científicas: características, funções e critérios de qualidade. In: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. (Org.). **Comunicação e produção científica**: contexto, indicadores, avaliação. São Paulo: Angellar, 2006. p. 163-190.

GONZALES AGUILAR, A.; RAMÍREZ POSADA, M.; FERREYRA, D. Tematres: software para gestionar tesauros. **El profesional de la información**, León, v. 21, n. 3, p. 319-325, mayo/jun. 2012.

HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14., Nantes, 1992. **Proceedings** [...] Nantes, 1992.

HEARST, M. A. Automated discovery of WordNet relations. In: FELLBAUM, C. (Ed.). **WordNet**: an electronic lexical database and some of its applications. Cambridge: MIT Press, 1998.

HEUSER, C. A. **Projeto de banco de dados**. 5. ed. Porto Alegre: Sagra Luzzatto, 2004.

HINZ, V. T. **Algoritmos para Interoperabilidade entre Ontologias.** 2008. 90 f. Dissertação (Mestrado em Informática) - Universidade Catolica de Pelotas, Pelotas, 2008.

HINZ, V. T.; PALAZZO, L. A. M. Algoritmos para Interoperabilidade entre Ontologias. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL. 2008, Rio de Janeiro. **Anais** [...] Rio de Janeiro, Universidade Federal Fluminense, 2008.

HJØRLAND, B. Theories are knowledge organizing systems (KOS). **Knowledge Organization**, Frankfurt, v. 42, n. 2, p. 113-128, 2015.

HODGE, G. Knowledge Organization Systems: an overview. In: HODGE, G. **Systems of knowledge organization for digital libraries**: beyond traditional authority files. Washington: Council on Library and Information Resources, 2000. p. 3-9.

IBEKWE-SAN JUAN, F.; BOWKER, G. C. Implications of big data for knowledge organization. **Knowledge Organization**, Frankfurt, v. 44, n. 3, p. 187-198, 2017.

IFLA - FEDERAÇÃO INTERNACIONAL DE ASSOCIAÇÕES DE BIBLIOTECAS E INSTITUIÇÕES. **Acesso e oportunidade para todos**: Como as bibliotecas contribuem para a agenda de 2030 das Nações Unidas. Netherlands: IFLA Headquarters, 2015. Disponível em: https://www.ifla.org/files/assets/hq/topics/libraries-development/documents/access-and-opportunity-for-all-pt.pdf. Acesso em: 05 ago. 2019.

INFORMAÇÃO & INFORMAÇÃO. 2019. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/index. Acesso em: 20 mar. 2019.

INFORMAÇÃO@PROFISSÕES. 2019. Disponível em: http://www.uel.br/revistas/uel/index.php/infoprof/index. Acesso em: 21 mar. 2019.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**. ISO 25964**: information and documentation - thesauri and interoperability with other vocabularies - part 1: thesauri for information retrieval. Genebra, 2011.

ISAAC, A.; SUMMERS, E. SKOS Simple Knowledge Organization System Primer. **Primer, World Wide Web Consortium (W3C)**, 2009. Disponível em: http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/. Acesso em: 12 maio 2019.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964:** information and documentation- thesauri and interoperability with other vocabularies - part 2: interoperability with other vocabularies. Genebra, 2013.

JESUS, D. L.; CUNHA, M. B. A biblioteca do futuro: um olhar no passado. **Informação & Informação**, Londrina, v. 24, n. 1, p. 1-30, mar. 2019.

JIA, J. ZHAO, J. Mapping Analysis of Pre-coordinated Classes in DDC and CLC. **Knowledge Organization**, Frankfurt, v. 42, n. 6, p. 369-385, 2015.

KALFOGLOU, Y.; SCHORLEMMER, M. Ontology mapping: the state of the art. **The knowledge engineering review**, Cambridge, v. 18, n. 1, p. 1-31, 2003.

KEMPF, A. O. RITZE, D. ECKERT, K. ZAPILKO, B. New Ways of Mapping Knowledge Organization Systems. **Knowledge Organization**, Frankfurt, v. 41, n. 1, p. 66-75, 2014.

KHOSRAVI, F.; VAZIFEDOOST, A. Creating a Persian ontology through thesaurus reengineering for organizing the Digital Library of the National Library of Iran. In: **International Conference on Libraries, Information and Society**. **Anais** [...] [S.l.], ICoLIS 2007. p. 19-36.

KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 1, p. A03-0, 2002. Disponível em: http://basessibi.c3sl.ufpr.br/brapci/v/a/1255. Acesso em: 01 ago. 2019.

LAIPELT, R. C. F. **Metodologia para seleção de termos equivalentes e descritores de tesauros:** um estudo no âmbito do Direito do Trabalho e do Direito Previdenciário. 2015. 213 f. Tese (Doutorado em Linguística Aplicada) - Universidade do Vale do Rio dos Sinos, São Leopoldo, 2015.

LANCASTER, F. W. **Construção e uso de tesauros**: curso condensado. Tradução de César Almeida de Menezes Silva. Brasília: Ibict, 1987. 114 p.

LANCASTER, F. W. **El control del vocabulario en la recuperación de información**. Tradução de Alejandro de La Cueva. Martín: Universitat de València, 1995.

LANCASTER, F. W. **El control del vocabulario en la recuperación de información**. Tradução de Alejandro de La Cueva. 2ª ed. Martín: Universitat de València, 2002.

LANCASTER, F. W.; SMITH, L. C. **Compatibility issues affecting information systems and services**. Paris: Unesco, 1983. 209 p.

LARA, M. L. G. Glossário: termos e conceitos da área de comunicação e produção científica. ln: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. **Comunicação e produção científica**: contexto, indicadores e avaliação. São Paulo: Algellara, 2006. p. 387-414.

MACHADO, P. N. **Extração de relações hiponímicas em corpora de língua portuguesa**. 80 f. 2015. Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2015.

MACHADO, P. N.; LIMA, V. L. S. Extração de relações hiponímicas em um *corpus* de língua portuguesa. **Revista de Estudos da Linguagem**, Belo Horizonte, v. 23, n. 3, p. 599-640, dez. 2015.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MILES, A.; BECHHOFER, S. SKOS simple knowledge organization system reference. **W3C recommendation**, v. 18, p. W3C, 2009. Disponível em: http://www.w3c.org/TR/2009/REC-skos-reference-20090818/.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. American Documentation, v.2, n.1, 1951, p. 20-32.

MOREIRA, M. P. **Ambiente para geração e manutenção semiautomática de tesauros**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

MOREIRA, W. Produção e leitura de hiperdocumentos: novos modos de interação leitor-texto. **Informação & Sociedade**, João Pessoa, v. 13, n. 1, 2003.

MOREIRA, W. Relações conceituais como ponto de inflexão entre linguagens documentais, terminologia e ontologias. **Scire**, Zaragoza, p. 123-127, 2012.

MOREIRA, W. SANTOS, J.C.F. VITORINI, E. F. Os padrões de hearst como recursos auxiliares semiautomáticos para a eficácia na leitura documentária. In: FUJITA, M. S. L.; NEVES, D. A. B.; DAL'EVEDOVE, P. R. (Org.). **Leitura documentária**: estudos avançados para a indexação. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2017. 318 p.

MOREIRO GONZÁLEZ, J. A. SÁNCHEZ CUADRADO, S. MORATO LARA, J. Mejora de la interoperabilidad semántica para la reutilización de contenidos mediante sistemas de organización del conocimiento. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 17, n. 33, p. 46-58, abr. 2012. ISSN 1518-2924.

MORIN, E. Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. **Traitement automatique des langues**, v. 40, n. 1, p. 143-166, 1999.

MUELLER, S. P. M. A ciência, o sistema de comunicação científica e a literatura científica. In: CAMPELLO, B. S.; CEDÓN, B. V.; KREMER, J. M. (Org.). **Fontes de informação para pesquisadores e profissionais.** Belo Horizonte: UFMG, 2000a.

MUELLER, S. P. M. O periódico científico. In: CAMPELLO, B. S.; CEDÓN, B. V.; KREMER, J. M. (Org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: UFMG, 2000b.

NEDELLEC, C.; NAZARENKO, A.; BOSSY, R. Information Extraction. In. STAAB, S.; STUDER, R. (Eds.). **Handbook on Ontologies**. 2 ed. New York: Springer Dordrecht Heidelberg London, 2019.

NEVILLE, H. H. Feasibility study of a scheme for reconciling thesauri covering a common subject. **Journal of documentation**, West Yorkshire, v. 26, n. 4, p. 313-336, 1970.

NEVILLE, H. H. Thesaurus reconciliation. **Aslib proceedings**, Londres, v.11, n. 24, p. 620-626,1972.

NOVELLINO, M. S. F. Instrumentos e metodologias de representação da informação. **Informação & Informação**, Londrina, v. 1, n. 2, p. 37-45, dez. 1996. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/1603. Acesso em: 26 jul. 2019.

OBRST, L. Ontologies for semantically interoperable systems. In: International conference on Information and knowledge management. 12, 2002. **Proceedings** [...] New Orleans: ACM, 2003. p. 366-369.

OSTIZ, H. C. **Descritores em ciências da saúde na área específica da fonoaudiologia brasileira**. 2010. 250 f. Tese (Doutorado em Ciência da Reabilitação) - Universidade de São Paulo, São Paulo, 2010.

OLIVEIRA, L. P. Linguística de *Corpus*: teoria, interfaces e aplicações. **Matraga-Revista do Programa de Pós-Graduação em Letras da UERJ**, Rio de Janeiro, v. 16, n. 24, 2009.

OLIVEIRA, R. R. **Recuperação Contextualizada de Documentos Integrados pelo Protocolo OAI-PMH**. 137 f. 2010. Dissertação (Mestrado em Ciência da Computação) — Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2010.

ONTOLOG. **Ontology Summit 2018 Communiqué**: Contexts in Context, 2018. Disponível em: http://ontologforum.org/index.php/OntologySummit2018. Acesso em: 08 ago. 2019.

ORENGO, V. M.; BURIOL, L.; COELHO, A. A study on the use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval. In: PETERS, C.; CLOUGH, P.; GEY, F. C.; KARLGREN, J.; MAGNINI, B.; OARD, B. W.; RIJKE, M.; STEMPFHUBER, M. (Eds.). **Evaluation of multilingual and multi-modal information retrieval**. Berlin: Springer, 2007. p. 91-98.

OUKSEL, A. M.; SHETH, A. Semantic interoperability in global information systems. **ACM Sigmod Record**, v. 28, n. 1, p. 5-12, 1999.

PASTOR-SÁNCHEZ, J. A.; MARTÍNEZ-MÉNDEZ, F. J. F. Manual de SKOS (simple knowledge organization system, sistema para la organización del conocimiento simple). **Anales de Documentación**, Murcia, v. 13, 2010. p. 285-320. Disponível em: http://skos.um.es/TR/skos-primer/. Acesso em: 02 ago. 2019.

PEDRO, G. W.; VALE, O. A. ComentCorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um *corpus* opinativo. In: FINATTO, M. J. B.; REBECHI, R. R.; SARMENTO, S.; BOCORNY, A. E. P. (Orgs.). **Linguística de *corpus*: perspectivas** [recurso eletrônico]. Porto Alegre: Instituto de Letras - UFRGS, 2018. 575 p.

PINHEIRO, L. V. R.; FERREZ, H. D**. Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: IBICT, 2014. Disponível em: http://www.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/tesauro-brasileiro-de-ciencia-da-informacao/resolveuid/ 1c60ede36f47aee60c48957ef6db7510. Acesso em: 12 maio 2019.

RAHM, E.; BERNSTEIN, P. A. A survey of approaches to automatic schema matching. **the VLDB Journal**, v. 10, n. 4, p. 334-350, 2001.

RAMALHO, R. A. S.; CERVANTES, B. M. N. Análise dos tipos de relações do modelo SKOS: perspectivas de representação de recursos audiovisuais. 2019 . In: CONGRESO ISKO ESPAÑA-PORTUGAL, 4, 2019. **Anais** [...] Barcelona: ISKO, 2019, (No Prelo).

RAMALHO, R. A. S. Ontologias e Knowledge Organization System (SKOS): aproximações e diferenças. In: GUIMARÃES, J. A. C.; DODEBEI, V. (Orgs.). **Organização do conhecimento e diversidade cultural**. Marília ISKO-Brasil: FUNDEPE, 2015. p. 100-107.

RODRIGUES, M. R.; PEREIRA, F. C. A.; LONDERO, R. R. ; CERVANTES, B. M. N. Tratamento temático da informação na revista Discursos Fotográficos: palavras-chave ou

descritores. In: SEMINÁRIO EM CIÊNCIA DA INFORMAÇÃO, 7, 2017, Londrina. **Anais** [...] Londrina: UEL, 2017, p. 1063-1076.

SANTOS, J. C. F. **Vocabulário controlado em periódicos científicos eletrônicos**: uma proposta de controle de termos. 144 f., 2015. Dissertação (Mestrado em Ciência da Informação) – Centro de Educação, Comunicação e Artes, Universidade Estadual de Londrina, Londrina, 2015.

SANTOS, J. C. F.; MOREIRA, W. SKOS: uma análise sobre as abordagens e suas as aplicações na Ciência da Informação. **Informação & Informação**, Londrina, v. 23, n. 3, p. 362-389, dez. 2018.

SANTOS, J. C. F. dos; CERVANTES, B.M. N.; FUJITA, M. S. L. Tesauro Eletrônico: importação no TemaTres e disponibilização na web. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina. **Anais**[...] Londrina: UEL; ANCIB, 2018.

SANTOS, J. C. F.; CERVANTES, B.M. N.; LONDERO, R. R. Controle de vocabulário: representação e acesso aos conteúdos temáticos de um periódico científico de comunicação. In: LOPES, M. I. V. de L.; RIBEIRO, N.; CASTRO, G. C. S.; BURNAY, C. D. (Org.). **XV Congresso IBERCOM 2017**: comunicação, diversidade e tolerância. São Paulo: ECA-USP; Lisboa: FCH-UCP, 2018. 6072 p.

SANTOS, J. C. F.; CERVANTES, B. M. N. Controle de vocabulário: palavras-chave como elemento representativo do conteúdo de publicações científicas no ambiente da Lusofonia Latino-Americana. In: CONFIBERCOM, 2., 2014, Braga. **Anais** [...] Minho: CECS-Centro de Estudos de Comunicação e Sociedade (Universidade do Minho), 2014.

SANTOS, J. C. F.; CERVANTES, B. M. N. Controle de vocabulário em periódicos científicos eletrônicos: proposta de compatibilização de palavras-chave. In: GUIMARÃES, J. A. C.; DODEBEI, V. (Org.). **Organização do conhecimento e diversidade cultural**. 1.ed. Marília: ISKO-Brasil; FUNDEPE, 2015a, v. 3, p. 262-271.

SANTOS, J. C. F.; CERVANTES, B.M. N. Controle de vocabulário em periódicos científicos eletrônicos: proposta de compatibilização de palavras-chave. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 16., 2015, João Pessoa. **Anais** [...] João Pessoa: UFPB; ANCIB, 2015b.

SANTOS, J. C. F. dos; FUJITA, M. S. L.; MOREIRA, W. Tesauro Unesp: Integração do registro de autoridade para o TemaTres. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 19., 2018, Londrina. **Anais** [...] Londrina: UEL; ANCIB, 2018.

SANTOS, J. C. F.; CERVANTES, B. M. N.; LONDERO, R. R.; GONCALEZ, P. R. V. A. Controle de vocabulário em periódicos científicos eletrônicos: proposta de implantação da VCPC *Tools* no periódico Discursos Fotográficos. In: SEMINÁRIO EM CIÊNCIA DA INFORMAÇÃO, 6., 2016, Londrina. **Anais** [...] Londrina: UEL, 2016.

SARDINHA, T. B. **Linguística de *Corpus***. Barueri, São Paulo: Manole, 2004. ISBN: 85-204-1676-4.

SAYÃO, L. F.; MARCONDES, C. H. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais. **Transinformação**, Campinas, v. 20, n. 2, p. 133-148, 2008.

SILVA, R. E.; SANTOS, P. L. V. A. C.; FERNEDA, E. Modelos de recuperação de informação e web semântica: a questão da relevância. **Informação & Informação**, Londrina, v. 18, n. 3, p. 27-44, out. 2013. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/12822. Acesso em: 01 ago. 2019.

SMIRAGLIA, R. P. **The Elements of Knowledge Organization**. Cham: Springer, 2014.

SOERGEL, D. **Indexing languages and thesauri:** Construction and maintenance. Los Angeles, CA: Melville, 1974. 632p.

SOERGEL, D. Unleashing the power of data through organization: structure and connections for meaning, learning and discovery. **Knowledge Organization**, Frankfurt, v. 42, n. 6, p. 401-427, 2015.

SOUZA, A. P. **Analisando conteúdos e mapeando informação em periódicos eletrônicos**: um estudo do periódico secundário PBCIB. 2011. 136 f. Dissertação (Mestrado em Ciência da informação) - Universidade Federal da Paraíba, João Pessoa, 2011.

SVENONIUS, Elaine. **The intellectual foundation of information organization**. Cambridge: MIT press, 2000.

TABA, L. S. **Extração automática de relações semânticas a partir de textos escritos em português do Brasil**. 98 f., 2013. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, 2013.

TABA, L. S.; CASELI, H. M. Automatic semantic relation extraction from Portuguese texts. In: LREC, 9., 2014. **Proceedings** [...] Reykjavik: ELRA, 2014. p. 2739-2746.

TEIXEIRA, J. R.; SOUZA, R. R. Conversão de tesauros em ontologias: um estudo exploratório. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 14., 2013. **Anais** [...] Florianópolis: UFSC, 2013.

TEMATRES. **TemaTres**: servidor de vocabulários controlados. 2018. Disponível em:http://r020.com.ar/tematres/manual/. Acesso em: 27 jul. 2018.

TORRES, C. E. A. **Uso de informação linguística e análise de conceitos formais no aprendizado de ontologias**. 2012. 87 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

UNISIST. **Study report on the feasibility of a world science information system**. Paris: UNESCO, 1971. 161 p.

WEITZEL, S. R. Fluxo de comunicação científica. ln: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. (org.). **Comunicação e produção científica**: contexto, indicadores e avaliação. São Paulo: Algellara, 2006. p. 81-114

YIN, R. K. **Estudo de caso**: planejamento e métodos. Tradução de Ana Thorell. Porto Alegre: Bookman, 2010.

LEI ZENG, Marcia; MAI CHAN, Lois. Trends and issues in establishing interoperability among knowledge organization systems. **Journal of the American Society for information science and technology**, v. 55, n. 5, p. 377-395, 2004.

ZENG, M. L.; HLAVA, M.; BUSCH, J. A.; BUCHEL, O.; ŽUMER, M. If you build it, will they come? a discussion of use cases and barriers of using the Knowledge Organization Systems (KOS) available as Linked Open Data (LOD). In: ASIS&T ANNUAL MEETING: INFORMATION SCIENCE WITH IMPACT: RESEARCH IN AND FOR THE COMMUNITY, 78., 2015. **Proceedings** [...] American Society for Information Science, 2015. p. 3.

APÊNDICES

APÊNDICE A - PADRÕES LÉXICOS-SINTÁTICOS ADAPTADOS/TRADUZIDOS PORTUGUÊS

Padrão Hearst (1992, 1998)	**Freitas (2007) Freitas e Quental (2007)**	**Baségio (2007) Torres (2012)**	**Taba (2013) Taba e Caseli (2014)**	**Machado (2015) Machado e Lima (2015)**	**Moreira, Santos e Vitorini (2017)**
NP_0 such as NP_1 {, NP_2…, (or | and) NP_i } for all NP_i, i ≥ 1, hyponym(NPi , NP0)	SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... , } (e | ou) SNi SN Hiper, (tais como | como_ PDEN) SN1 { , SN2 ... ,}(e | ou) SNi	SUB como {(SUB,)*(ou|e) } SUB SUB tal(is) como {(SUB,)*(ou|e) } SUB	SN_Hiper (tais como | como) SN {, SN} *(e|ou) SN	SN(,)? como (SN ,)*(SN (e|ou))*SN	SN (tais como | como) SN { , SN ... , } (e | ou) SN
such NP as {NP, }*{(or|and)} NP	não utilizado	tal(is) SUB como {(SUB,)*(ou|e) } SUB	não utilizado	SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN	tal(is) SN como {(SN,)*(ou|e)} SN
NP {, NP}* {,} or other NP	SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper	SUB {, SUB}* {,} ou outro(s) SUB	SN {, SN}* , ? (e|ou) outros SN_Hiper	SN {, SN}* , ? (e|ou) outros SN_Hiper	SN {, SN}* {,} ou outro(s) SN
NP {, NP}* {,} and other NP		SUB {, SUB}* {,} e outro(s) SUB			SN {, SN}* {,} e outro(s) SN
NP {,} including {NP,}*{or|and} NP	não utilizado	SUB {,} incluindo {SUB,}*{ou|e} SUB	não utilizado	SN(,)? incluindo (SN ,)*(SN (e|ou))*SN	SN {,} incluindo {SN,}*{ou|e} SN
NP {,} especially {NP,}*{or|and} NP	não utilizado	SUB {,} especialmente {SUB,}*{ou|e} SUB SUB {,} principalmente {SUB,}*{ou|e} SUB SUB {,} particularmente {SUB,}*{ou|e} SUB SUB {,} em especial { SUB,}*{ou|e}	não utilizado	SN(,)? especialmente (SN ,)*(SN (e|ou))*SN	SN {,} especialmente {SN,}*{ou|e} SN (6b) SN {,} principalmente {SN,}*{ou|e} SN (6c) SN {,} particularmente {SN,}*{ou|e} SN (6d) SN {,} em especial { SN,}*{ou|e} SN (6e) SN {,} em particular { SN,}*{ou|e} SN

		SUB SUB {,} em particular { SUB,}*{ou|e} SUB SUB {,} de maneira especial { SUB,}*{ou|e} SUB SUB {,} sobretudo { SUB,}*{ou|e} SUB			(6f) SN {,} de maneira especial { SN,}*{ou|e} SN (6g) SN {,} sobretudo { SN,}*{ou|e} SN
-	tipos de SN Hiper: SN1{ , SN2... ,} (e | ou) SNi	-	tipos de SN_Hiper: SN {, SN}* (e|ou) SN	<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN	-
-	SN HHiper chamado/s/a/as (de) SN Hipo	-	SN_Hiper chamad(o|a|os|as) de? SN	SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN	-
-	*SN Hiper conhecido/s/a/as como SN Hipo.	-	-	SN((,)? também)?(,|é|são |foram)? conhecid(o|a|os| as) como (SN ,)*SN (e|ou))*SN"	-
-	-	-	SN {,SN} * ,? (e|ou) (qualquer|quais quer) outro{s}? SN_Hiper	(SN (ou|e|,))*< (qualquer|quaisq uer) outr(a|o)(s)? sn>	-
-	-	-	SN é (o|a|um|uma) SN_Hiper	SN é < (o|a) sn> SN é < (um|uma) sn>	-
-	-	-	SN são SN_Hiper	SN são SN	-

Fonte: Elaborado pelo autor (2020) adaptado de Moreira, Santos e Vitorini (2017)

APÊNDICE B - ANÁLISE DOS ARTIGOS DA CATEGORIA "APLICAÇÃO – VOCABULÁRIOS"

AUTOR	OBJETO	LOCAL AMBIÊNCIA	AGENTE	CAUSA E EFEITO
Albertoni, De Martino, Di Franco, De Santis & Plini (2014)	Tesauro - Dados Ambientais	*EARTh Linked Data*	CNR-IIA-EKOLab *CNR-IMATI*.	Conexão de diferentes tesauros da área.
Ameri, Kulvatunyou, Ivezic & Kaikhah (2014)	Tesauro – Domínio Industrial	ManuTerms	-	Conceitualização ontológica[12]
Balkan & Bell (2014)	Tesauros áreas Ciência Social e Humanas	Humanities and Social Science Electronic Thesaurus (HASSET) European Language Social Science Thesaurus (ELSST)	UK Data Archive	Conversão dos tesauros em um modelo baseado em conceito, com uma versão SKOS
Binding & Tudhope (2016)	Vocabulários arqueológicos de diferentes países.	Getty Vocabulary LOD SPARQL	Projeto ARIADNE	Mapeamento de vocabulários
Çağdaş & Stubkjær (2015)	Tesauro para o domínio de cadastro e administração da terra	Linked Land Administration	-	Construção de tesauro em SKOS
Caracciolo et al. (2012)	Tesauro de Agricultura	AGROVOC	Food and Agriculture Organisation (FAO)	Alinhamento com outro SOC, abordagem de

[12] A conceitualização ontológica refere-se ao processo de criação de uma visão abstrata do domínio de interesse, por meio de um conjunto de conceitos interconectados (AMERI, KULVATUNYOU; IVEZIC; KAIKHAH, 2014).

				direções atuais e futuras
Cohen & Franke (2015)	Vocabulário Militar	C2 ASCA DART	-	Interoperabilidade entre os vocabulários
Firmino & Baptista (2013)	Vocabulário controlado Comunicação	-	Autoridade Nacional de Comunicações (ANACOM)	Organização de termos de descrição de recursos e desenvolvimento do vocabulário controlado (SKOS – LOD)
Gray, Gray, Hall & Ounis (2010)	Vocabulário Controlado Astronomia	Vocabulary Explorer Vocabulary Explorer Web	Virtual Observatory	Recuperação de recursos por meio de vocabulários
Hubain, De Wilde & Van Hooland (2016)	Tesauro Indústria biofarmacêutica	International Society for Pharmaceutical Engineering Glossary	Biotech Quality Group	Concepção automática de tesauro multilingue
Jia & Wei (2012)	Tesauro	Chinese Thesaurus (CT) Library of Congress Subject Headings (LCSH)	Scientific and Technical Information Institute of China – National Library of China	Mapeamento entre SOC
Ma Luisa Alvite, Pérez-León, MartÍnez-González, & Dámaso-Javier Vicente (2010)	Tesauro Multidisciplinar	Eurovoc	-	Proposta de integração do tesauro com sistemas de informação jurídica

Ma, Carranza, Wu, Van Der Meer & Liu (2011)	Tesauro Multilingue Geológico	Geological Time Scale (GTS) Multilingual Thesaurus of Geological Time Scale (MLTGTS)	-	Superar barreiras linguísticas na integração com mapas on-line
Martínez-González & Alvite Díez (2014)	Ferramentas Tesauro	-	-	Framework para avaliação de ferramentas de tesauro
Méndez & Greenberg (2012)	SOC	Helping Interdisciplinary Vocabulary Engineering (HIVE)	-	Abordagem de tecnologias aplicadas SOC (LOD, SKOS)
Miles & Pérez-Agüera (2007)	Taxonomia Tesauro Sistema de Classificação	Morten Frederiksen's Web log categories *UK Archival Thesaurus* Physics and Astronomy Classification Scheme (PACS)	-	Exemplos de aplicação prática do SKOS
Nicholson & McCulloch (2006)	Descrição de assunto	Project HILT	Joint Information Systems Committee (JISC)	Mapeamento diferentes assuntos e esquemas de classificação
O'DELL (2015)	Vocabulário Ilustrado	Artists' Books Thesaurus	-	Integração de imagens para um tesauro - visual
Papadakis & Kyprianos (2011)	Tesauro para sistema IR	Library of Congress Subject Headings (LCSH)	-	Mapeamento LCSH com tesauro para Integração em sistema IR
Pastor Sánchez (2013a)	Marcação Semântica	Tesauro Unesco	-	Marcação semântica a partir da web semântica e LOD

Pastor-Sanchez (2013b)	Dados abertos	Mads/RDF e GND	-	Controle de Autoridade conectado com vocabulários controlados e modelos de descrição
Pastor-Sánchez, Martínez-Méndez & Rodríguez-Muñoz (2012)	Controle de vocabulário Tesauro	The data hub	-	Análise tipologia de vocabulários
Sanchez-Alonso & Garcia-Barriocanal (2006)	Interoperabilidade de Esquemas conceituais	OpenCyc	-	Mapeamento dos metadados do SKOS para um modelo baseado em ontologias
Singthongchai, Niwattanakul & Chamnongsri (2016)	SOC Agricultura	Thai Agricultural Thesaurus	-	Técnica de ponderação de termos com coeficiente de correlação
Suominen & Mader (2014)	Vocabulário Controlado	Data Hub qSOKS	-	Avaliação e conjunto de heurísticas para correção automática de problemas
Tennis & Sutton (2008)	Vocabulário Controlado	Dublin Core Metadata Initiative (DCMI)	-	Modelagem SKOS — instância do conceito - para mediar entre o SKOS conceito abstrato e o esquema concreto

Wright, Harrison & Watkins (2015)	Tesauro dados de química ambiental	CEH Analytical Services Thesaurus (CAST)	Laboratory Information Management System (LIMS)	Interoperabilidade semântica entre os recursos etiquetados, mapeamento de demais SOC
Zapilko, Schaible, Mayr & Mathiak (2013)	Tesauro Ciências Sociais	Thesaurus for the Social Sciences (TheSoz)	Leibniz Institute for the Social Sciences (GESIS)	Mapeamento do tesauro em SKOS - LOD

Fonte: Santos e Moreira (2018).

APÊNDICE C – AMOSTRAGEM DO MAPEAMENTO DE PALAVRAS-CHAVE REVISTA INFORMAÇÃO & INFORMAÇÃO

Palavra-chave	Termo TBCI	Níveis de correspondência (Neville, 1970)	Mapeamento
AACR2	AACR2	Caso 1	IDENTICO
Abordagem Sociocultural da Ciência da Informação	ciência da informação	Caso 4	I_INDICE
acepção	análise linguística	Caso 4	CORRESP
Acervo Bibliográfico	acervos bibliográficos	Caso 1	IDENTsPL
Acesso	acesso	Caso 1	IDENTICO
Acesso à informação	acesso à informação	Caso 1	IDENTICO
Acesso Aberto	acesso aberto	Caso 1	IDENTICO
Acesso Aberto à literatura científica	acesso aberto	Caso 4	I_INDICE
Administração	administradores	Caso 1	IDENTsSUF
Agentes de interface	usabilidade (programas de computador)	Caso 4	CORRESP
Banco de dados bibliográficos	bases de dados bibliográficos	Caso 2	CORRESP
Bancos	bases de dados	Caso 2	CORRESP
Base de Dados	bases de dados	Caso 1	IDENTsPL
Bases de Dados	bases de dados	Caso 1	IDENTICO
Bebeteca	bibliotecas infantis	Caso 2	CORRESP
Benchmarking	benchmarking	Caso 1	IDENTICO
Bibliografia	bibliografias	Caso 1	IDENTsPL
Bibliografia - história e teoria	bibliografias	Caso 4	CORRESP
Bibliografia e memória	bibliografias	Caso 4	CORRESP
Bibliografia material	bibliografias	Caso 4	CORRESP
Campos e Disciplinas	campos e disciplinas	Caso 1	IDENTICO
Canais de comunicação científica	canais de comunicação	Caso 4	I_INDICE
Canais de informação – Universidade Estadual de Ponta Grossa	canais de comunicação	Caso 4	CORRESP
Cartas patrimoniais	cartas	Caso 4	I_INDICE
Catalogação	catalogação	Caso 1	IDENTICO
Catalogação automatizada	catalogação automatizada	Caso 1	IDENTICO
Catalogação descritiva	catalogação descritiva	Caso 1	IDENTICO
Catálogo on-line	catálogos	Caso 4	CORRESP
Catálogos	catálogos	Caso 1	IDENTICO
Categorização	categorias	Caso 1	IDENTsSUF
Dados científicos	dados científicos	Caso 1	IDENTICO
Dados de pesquisa	dados de pesquisa	Caso 1	IDENTICO
Definição de metadados	metadados	Caso 4	CORRESP
demanda de informação	demanda de informação	Caso 1	IDENTICO
Desenvolvimento de Coleção	desenvolvimento de coleções	Caso 1	IDENTsPL
Digitalização	digitalização	Caso 1	IDENTICO

Direito à informação	direito à informação	Caso 1	IDENTICO
Disseminação da informação	disseminação da informação	Caso 1	IDENTICO
Disseminação Seletiva da Informação	disseminação seletiva da informação	Caso 1	IDENTICO
Dissertações - Resumos	dissertações e teses	Caso 2	CORRESP
E-book	e-books	Caso 1	IDENTsPL
Ecologia/Meio Ambiente	ciências biológicas	Caso 4	CORRESP
Economia	economia	Caso 1	IDENTICO
Economia da Informação	economia da informação	Caso 1	IDENTICO
Economia Política	economia	Caso 4	I_INDICE
Economia Política da Comunicação	políticas de informação	Caso 4	CORRESP
Edição de periódicos	editoração científica	Caso 4	CORRESP
Educação	educação	Caso 1	IDENTICO
educação a distância	educação a distância	Caso 1	IDENTICO
Educação a Distância - Bibliotecas Híbridas	educação a distância	Caso 4	I_INDICE
Faculdade de Biblioteconomia	faculdades de biblioteconomia	Caso 1	IDENTsPL
Fatores críticos de sucesso	fatores críticos de sucesso	Caso 1	IDENTICO
Fluxo de informação	fluxo da informação	Caso 1	IDENTsPV
Fluxos de Informação	fluxo da informação	Caso 1	IDENTsPL
Folksonomia	folksonomias	Caso 1	IDENTsPL
Folksonomias	folksonomias	Caso 1	IDENTICO
Formação do Profissional Bibliotecário	formação profissional	Caso 4	I_INDICE
Formação Profissional	formação profissional	Caso 1	IDENTICO
Formato MARC	formatos MARC	Caso 1	IDENTsPL
Formato MARC 21 para Dados de Autoridade	formatos MARC	Caso 4	CORRESP
Gerenciamento Documental	gerenciamento	Caso 4	I_INDICE
Gestão	gestão	Caso 1	IDENTICO
Gestão ambiental	gestão	Caso 4	I_INDICE
Gestão da Coleção	gestão de coleções	Caso 1	IDENTsPL
Gestão da Informação	gestão da informação	Caso 1	IDENTICO
Gestão da Informação – formação profissional	gestão da informação	Caso 4	I_INDICE
Gestão da segurança da informação	gestão da informação	Caso 4	I_INDICE
Gestão de Biblioteca	gestão de bibliotecas	Caso 1	IDENTsPL
Gestão de bibliotecas	gestão de bibliotecas	Caso 1	IDENTICO
Gestão de dados de pesquisa	dados de pesquisa	Caso 4	I_INDICE
Hábito	hábitos de coleta da informação	Caso 2	CORRESP
Hipertexto	hipertextos	Caso 1	IDENTsPL
História	história	Caso 1	IDENTICO
História da ciência	história da ciência da informação	Caso 2	CORRESP
História da Ciência da Informação	história da ciência da informação	Caso 1	IDENTICO
História da Organização do Conhecimento	organização do conhecimento	Caso 4	I_INDICE
História do Livro e das Bibliotecas	história das bibliotecas	Caso 4	I_INDICE

história oral	história	Caso 4	I_INDICE
Imagem	imagens	Caso 1	IDENTsPL
Imagens Técnicas	imagens	Caso 4	I_INDICE
Inclusão digital	inclusão digital	Caso 1	IDENTICO
Inclusão social	inclusão social	Caso 1	IDENTICO
Indexação	indexação	Caso 1	IDENTICO
Indexação de assunto	indexação de assuntos	Caso 1	IDENTsPL
Indexação de Imagens	indexação de imagens	Caso 1	IDENTICO
Indexação de imagens em movimento	indexação de imagens	Caso 4	I_INDICE
Indexação social	indexação	Caso 4	I_INDICE
Indicadores científicos	indicadores	Caso 4	I_INDICE
Jornalismo científico	jornalismo científico	Caso 1	IDENTICO
Lei de Acesso à Informação	acesso à informação	Caso 4	I_INDICE
leitura	promoção da leitura	Caso 2	CORRESP
Letramento informacional	letramento informacional	Caso 1	IDENTICO
Liberdade Intelectual	liberdade intelectual	Caso 1	IDENTICO
linguagem	barreiras de linguagem	Caso 2	CORRESP
Linguagem documentária	linguagens documentárias	Caso 1	IDENTsPL
Linguagens de indexação	linguagens de indexação	Caso 1	IDENTICO
Linguagens de marcação	linguagens de marcação	Caso 1	IDENTICO
Linguagens documentárias	linguagens documentárias	Caso 1	IDENTICO
Linguística documentária	linguística	Caso 4	I_INDICE
Mapas conceituais	mapas	Caso 4	I_INDICE
Marketing	marketing	Caso 1	IDENTICO
Marketing de relacionamento	marketing	Caso 4	CORRESP
Mecanismos de Busca	mecanismos de busca	Caso 1	IDENTICO
Mediação da informação	mediadores da informação	Caso 1	IDENTsSUF
Medicina	medicina	Caso 1	IDENTICO
Mercado de Trabalho	mercado de trabalho	Caso 1	IDENTICO
Mercadoria	informação para negócios	Caso 2	CORRESP
Metadados	metadados	Caso 1	IDENTICO
Metodologia	métodos de pesquisa	Caso 4	CORRESP
Necessidade de Informação	necessidades de informação	Caso 1	IDENTsPL
Necessidades de informação de Avicultores	necessidades de informação	Caso 4	I_INDICE
Necessidades informacionais – Universidade Estadual de Ponta Grossa	necessidades de informação	Caso 2	CORRESP
Normalização	normalização	Caso 1	IDENTICO
Normas técnicas	normalização	Caso 4	CORRESP
NTICs – Educação a Distância	educação a distância	Caso 4	I_INDICE
objeto de estudo	objetos digitais	Caso 2	CORRESP
Objeto digital	objetos digitais	Caso 1	IDENTsPL
Odontologia	odontologia	Caso 1	IDENTICO
Ontologia	ontologias	Caso 1	IDENTsPL

Ontologia semiótica	ontologias	Caso 4	CORRESP
Ontologias	ontologias	Caso 1	IDENTICO
Organização da Informação	organização da informação	Caso 1	IDENTICO
Organização de aprendizagem	organização do conhecimento	Caso 2	CORRESP
Organização do conhecimento	organização do conhecimento	Caso 1	IDENTICO
Organização e representação do conhecimento	organização do conhecimento	Caso 4	I_INDICE
Palavras-chaves: Ética profissional	listas de palavras não-significativas	Caso 2	CORRESP
Paradigma	paradigmas	Caso 1	IDENTsPL
Paradigma cognitivo	paradigmas cognitivos	Caso 1	IDENTsPL
Paradigma indiciário	paradigmas cognitivos	Caso 2	CORRESP
Paradigma social	paradigmas sociais	Caso 1	IDENTsPL
Paradigmas culturais	paradigmas	Caso 4	I_INDICE
Paradigmas da Informação	paradigmas cognitivos	Caso 2	CORRESP
Patentes	patentes	Caso 1	IDENTICO
Pensamento Crítico	liberdade de pensamento	Caso 2	CORRESP
Pequenas e médias empresas	pequenas e médias empresas	Caso 1	IDENTICO
RDF	RDF	Caso 1	IDENTICO
Recuperação da informação	recuperação da informação	Caso 1	IDENTICO
Recuperação de informações	recuperação da informação	Caso 1	IDENTsPL
Recursos informacionais	recursos informacionais	Caso 1	IDENTICO
Rede de computador	redes de computadores	Caso 1	IDENTsPL
Rede de Informação	redes de informação	Caso 1	IDENTsPL
Rede Social	redes sociais	Caso 1	IDENTsPL
Redes de Comunicação de Dados	redes de comunicação	Caso 4	I_INDICE
Redes de Informação	redes de informação	Caso 1	IDENTICO
Redes Eletrônicas de Comunicação	redes de comunicação	Caso 4	I_INDICE
Saúde	ciências da saúde	Caso 2	CORRESP
Semântica	semântica	Caso 1	IDENTICO
Semiótica	semiótica	Caso 1	IDENTICO
Serviço de informação	serviços de informação	Caso 1	IDENTsPL
Serviço de Referência	serviços de referência	Caso 1	IDENTsPL
Serviço de referência digital	serviços de referência	Caso 4	CORRESP
Serviço de Referência Virtual	serviços de referência	Caso 4	CORRESP
Serviços de Informação	serviços de informação	Caso 1	IDENTICO
Serviços e Produtos de informação	serviços de informação	Caso 4	I_INDICE
Serviços em Unidades de Informação	serviços de informação	Caso 4	I_INDICE
Taxonomia	taxonomias	Caso 1	IDENTsPL
Tecnologia da Informação	tecnologias da informação	Caso 1	IDENTsPL
Tecnologia da Informação e Comunicação	tecnologias da informação e comunicação	Caso 1	IDENTsPL
Tecnologia de Informação	tecnologias da informação	Caso 1	IDENTsPL
Tecnologia de Informação e Comunicação	tecnologias da informação e comunicação	Caso 1	IDENTsPL
Tecnologias da Informação	tecnologias da informação	Caso 1	IDENTICO

Tecnologias da Informação e Comunicação	tecnologias da informação e comunicação	Caso 1	IDENTICO
Tecnologias da Informação e da Comunicação	tecnologias da informação e comunicação	Caso 1	IDENTsPV
Tecnologias de informação	tecnologias da informação	Caso 1	IDENTsPV
Tecnologias de Informação e Comunicação	tecnologias da informação e comunicação	Caso 1	IDENTsPV
Unidades de Informação	unidades de informação	Caso 1	IDENTICO
Universidade	universidades	Caso 1	IDENTsPL
Usabilidade	usabilidade (programas de computador)	Caso 1	CORRESP
Usuário da Informação	usuários de informação	Caso 1	IDENTsPL
Vocabulário controlado	vocabulários controlados	Caso 1	IDENTsPL
Web	Web	Caso 1	IDENTICO
Web 2.0	Web	Caso 4	I_INDICE
Web Semântica	web semântica	Caso 1	IDENTICO
Web Sites	Web	Caso 4	I_INDICE
Web Social	Web	Caso 4	I_INDICE
Web Social Semântica	web semântica	Caso 4	I_INDICE
Website	sítios web	Caso 2	CORRESP
Websites	sítios web	Caso 2	CORRESP
World Wide Web	World Wide Web	Caso 1	IDENTICO
XML	XML	Caso 1	CORRESP

Fonte: Dados da pesquisa.

APÊNDICE D - MAPEAMENTO DE PALAVRAS-CHAVE REVISTA INFORMAÇÃO@PROFISSÕES

Palavra-chave	Termo TBCI	Níveis de correspondência (Neville, 1970)	Mapeamento
Acessibilidade	acessibilidade	Caso 1	IDENTICO
Administração	administradores	Caso 1	IDENTsSUF
Arquitetura da Informação	arquitetura de informação	Caso 1	IDENTsPV
Arquivista	arquivistas	Caso 1	IDENTsPL
Arquivologia	arquivologia	Caso 1	IDENTICO
Auditoria interna	auditoria	Caso 4	I_INDICE
bibliometria	bibliometria	Caso 1	IDENTICO
Biblioteca	bibliotecas	Caso 1	IDENTsPL
Biblioteca Especializada	bibliotecas especializadas	Caso 1	IDENTsPL
Biblioteca Pública	bibliotecas públicas	Caso 1	IDENTsPL
Bibliotecas escolares	bibliotecas escolares	Caso 1	IDENTICO
Bibliotecas universitárias	bibliotecas universitárias	Caso 1	IDENTICO
Biblioteconomia	biblioteconomia	Caso 1	IDENTICO
Busca	buscas	Caso 1	IDENTsPL
Catalogação	catalogação	Caso 1	IDENTICO
Centros de informação	centrais de informação	Caso 1	IDENTsSUF
Ciência da Informação	ciência da informação	Caso 1	IDENTICO
Ciência da Informação - Biblioteconomia	ciência da informação	Caso 4	I_INDICE
Ciências Sociais Aplicadas/Ciência da Informação	ciência da informação	Caso 4	I_INDICE
cocitações	cocitação	Caso 1	IDENTsPL
Competência em Informação	competência em informação	Caso 1	IDENTICO
Competência informacional	competência informacional	Caso 1	IDENTICO
Competência Profissional	competências profissionais	Caso 1	IDENTsPL
Comunidades de prática	comunidades de prática	Caso 1	IDENTICO
Cultura organizacional	cultura organizacional	Caso 1	IDENTICO
Desenvolvimento de Coleções	desenvolvimento de coleções	Caso 1	IDENTICO
Direito	direito	Caso 1	IDENTICO
Educação	educação	Caso 1	IDENTICO
Educação de Usuário	educação de usuários	Caso 1	IDENTsPL
Estratégia de busca	estratégias de busca	Caso 1	IDENTsPL
Folksonomia	folksonomias	Caso 1	IDENTsPL
Gestão de Documentos	gestão de documentos	Caso 1	IDENTICO
Gestão do conhecimento	gestão do conhecimento	Caso 1	IDENTICO
História	história	Caso 1	IDENTICO
Information Literacy	information literacy	Caso 1	IDENTICO
Interdisciplinaridade	interdisciplinaridade	Caso 1	IDENTICO
Interoperabilidade semântica	interoperabilidade	Caso 4	I_INDICE
Linguagens de Marcação	linguagens de marcação	Caso 1	IDENTICO
Livro	livros	Caso 1	IDENTsPL

Livros eletrônicos	livros eletrônicos	Caso 1	IDENTICO
Mecanismos de busca	mecanismos de busca	Caso 1	IDENTICO
Mediação da Informação	mediadores da informação	Caso 1	IDENTsSUF
Museologia	museologia	Caso 1	IDENTICO
Organização da informação	organização da informação	Caso 1	IDENTICO
Organização do Conhecimento	organização do conhecimento	Caso 1	IDENTICO
Organização temática da informação	organização da informação	Caso 4	I_INDICE
Pesquisa	pesquisa	Caso 1	IDENTICO
Pesquisa escolar	pesquisa	Caso 4	I_INDICE
Pesquisa significativa	pesquisa	Caso 4	I_INDICE
Política de aquisição	políticas de aquisição	Caso 1	IDENTsPL
Política de desenvolvimento de coleções	políticas de desenvolvimento de coleções	Caso 1	IDENTsPL
Política de Informação	políticas de informação	Caso 1	IDENTsPL
Preservação Digital	preservação digital	Caso 1	IDENTICO
Profissional da Informação	profissionais de informação	Caso 1	IDENTsPL
Projetos de pesquisa	projetos de pesquisa	Caso 1	IDENTICO
Psicologia	psicologia	Caso 1	IDENTICO
Redes de bibliotecas	redes de bibliotecas	Caso 1	IDENTICO
Registro bibliográfico	registros bibliográficos	Caso 1	IDENTsPL
Repositórios Institucionais	repositórios institucionais	Caso 1	IDENTICO
Representação da Informação	representação da informação	Caso 1	IDENTICO
Segurança da Informação	segurança da informação	Caso 1	IDENTICO
Sistemas de Informações	sistemas de informação	Caso 1	IDENTsPL
Tecnologias da Informação	tecnologias da informação	Caso 1	IDENTICO
Teologia	teologia	Caso 1	IDENTICO
Web Semântica	web semântica	Caso 1	IDENTICO

Fonte: Dados da pesquisa.

APÊNDICE E - MAPEAMENTO DE PALAVRAS-CHAVE REVISTA DISCURSOS FOTOGRÁFICOS

Palavra-chave	Termo Comunicação Visual	Níveis de correspondência (Neville, 1970)	Mapeamento
Acervos Cinematográficos	Acervos cinematográficos	Caso 1	IDENTICO
Acervos Fotográficos	Acervos fotográficos	Caso 1	IDENTICO
Análise do Discurso	Análise do discurso	Caso 1	IDENTICO
Análise do Discurso	Análise do discurso	Caso 4	I_INDICE
Análise Fílmica	Análise fílmica	Caso 1	IDENTICO
Análise Fotográfica	Análise fotográfica	Caso 1	IDENTICO
Análise Publicitária	Análise publicitária	Caso 1	IDENTICO
Análise Semiótica	Análise semiótica	Caso 1	IDENTICO
Análise Televisiva	Análise televisiva	Caso 1	IDENTICO
Análise Videográfica	Análise videográfica	Caso 1	IDENTICO
Animações	Animações	Caso 1	IDENTICO
Antropologia Visual	Antropologia visual	Caso 1	IDENTICO
Artes	Artes	Caso 1	IDENTICO
Assessoria de Imprensa	Assessoria de imprensa	Caso 1	IDENTICO
Caricatura	Caricatura	Caso 1	IDENTICO
Cartum	Cartum	Caso 1	IDENTICO
Charge	Charges	Caso 1	IDENTsPL
Cinejornal	Cinejornal	Caso 1	IDENTICO
Cinema Regional	Cinema	Caso 4	I_INDICE
Comunicação Mercadológica	Comunicação mercadológica	Caso 1	IDENTICO
Cultura Visual	Cultura visual	Caso 1	IDENTICO
Design de Moda	Design de moda	Caso 1	IDENTICO
Design Gráfico	Design gráfico	Caso 1	IDENTICO
Design Industrial	Design	Caso 4	I_INDICE
Diagramação	Diagramação	Caso 1	IDENTICO
Documentário	Documentário	Caso 1	IDENTICO
Documentário imaginário	Documentário	Caso 4	I_INDICE
Edição Jornalística	Edição jornalística	Caso 1	IDENTICO
Etnografia	Etnografia	Caso 1	IDENTICO
Filosofia	Filosofia	Caso 1	IDENTICO
Filosofia da Linguagem	Filosofia da linguagem	Caso 1	IDENTICO
Fotoclube	Fotoclubismo	Caso 1	IDENTsSUF
Fotoclubismo	Fotoclubismo	Caso 1	IDENTICO
Fotoetnografia	Fotoetnografia	Caso 1	IDENTICO
Fotografia	Fotografia	Caso 1	IDENTICO
Fotografia Analógica	Fotografia analógica	Caso 1	IDENTICO
Fotografia Contemporânea	Fotografia	Caso 4	I_INDICE
Fotografia Digital	Fotografia digital	Caso 1	IDENTICO
Fotografia Documental	Fotografia documental	Caso 1	IDENTICO
Fotografia e Legislação	Fotografia e legislação	Caso 1	IDENTICO

Fotografia e memória	Fotografia e memória	Caso 1	IDENTICO
Fotografia Erótica	Fotografia	Caso 4	I_INDICE
Fotografia menor	Fotografia	Caso 4	I_INDICE
fotografia participativa	Fotografia	Caso 4	I_INDICE
Fotografia Publicitária	Fotografia	Caso 4	I_INDICE
Fotografia Urbana	Fotografia	Caso 4	I_INDICE
Fotojornalismo	Fotojornalismo	Caso 1	IDENTICO
História da Fotografia	História da fotografia	Caso 1	IDENTICO
História da teoria da fotografia	História da fotografia	Caso 4	I_INDICE
Ilustração	Ilustração	Caso 1	IDENTICO
Infografia	Infografia	Caso 1	IDENTICO
Jornalismo Digital	Jornalismo	Caso 4	I_INDICE
Jornalismo Esportivo	Jornalismo	Caso 4	I_INDICE
Jornalismo Impresso	Jornalismo	Caso 4	I_INDICE
Jornalismo Literário	Jornalismo	Caso 4	I_INDICE
Jornalismo moderno	Jornalismo	Caso 4	I_INDICE
Jornalismo online	Jornalismo	Caso 4	I_INDICE
Jornalismo Popular	Jornalismo	Caso 4	I_INDICE
Minissérie	Minissérie	Caso 1	IDENTICO
Revistas	Revistas	Caso 1	IDENTICO
Revistas Jornalísticas	Revistas	Caso 4	I_INDICE
Semiótica cultural	Semiótica da cultura	Caso 1	IDENTsSUF
Semiótica da Cultura	Semiótica da cultura	Caso 1	IDENTICO
Semiótica Discursiva	Semiótica	Caso 4	I_INDICE
Semiótica Francesa	Semiótica	Caso 4	I_INDICE
Técnicas de Fotografia	Técnicas de fotografia	Caso 1	IDENTICO
Técnicas de Fotografia	Técnicas de fotografia	Caso 1	IDENTICO
Telejornalismo	Telejornalismo	Caso 1	IDENTICO
Teleteatro	Teleteatro	Caso 1	IDENTICO
televisão	Televisão	Caso 1	IDENTICO
Teoria Semiótica	Teoria semiótica	Caso 1	IDENTICO
Teorias da Fotografia	Teorias da fotografia	Caso 1	IDENTICO
Vídeo Comunitário	Vídeo	Caso 4	I_INDICE
Vídeo Documentário	Vídeo documentário	Caso 1	IDENTICO
Vídeo Jockey	Vídeo	Caso 4	I_INDICE

Fonte: Dados da pesquisa.