XVII Encontro Nacional de Pesquisa em Ciência da Informação **(XVII ENANCIB)**

GT 8 – Informação e Tecnologia

ONTOLOGIAS NO PROCESSO DE INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS

ONTOLOGIES IN AUTOMATIC INDEXING PROCCESS OF TEXTUAL DOCUMENTS

Eder Antonio Pansani Junior[1], Edberto Ferneda[2]

Modalidade da apresentação: Comunicação Oral

Resumo: A evolução da tecnologia e do acesso por meio da internet mudou a forma como a sociedade lida com documentos e informações. A possibilidade de armazenar e compartilhar o conhecimento registrado em formato digital expande os horizontes da ciência e da pesquisa. O processo de recuperação de informações estuda a relação entre o acervo documental e a representação temática dos registros que o compõem e os usuários e suas necessidades informacionais, que se traduzem em expressões elaboradas utilizando uma linguagem. Um sistema de recuperação de informação é um elemento mediador entre um acervo documental e seus requisitantes. Um dos aspectos que interfere diretamente na sua eficiência é a forma como os documentos são representados. Sendo assim, pesquisas sobre indexação automática tomam importância, principalmente em ambiente de grande produção e disseminação de documentos, como é o caso da Web. A utilização de vocabulários controlados como elementos de normalização terminológica é um recurso utilizado para melhorar os resultados do processo de indexação. Este trabalho tem por objetivo propor e desenvolver um método de utilização de ontologias no processo de indexação automática de documentos textuais, fazendo uso da estrutura lógica e conceitual das ontologias de domínio e implementado um método que permite aos sistemas de indexação automática a realização de inferências automáticas, favorecendo uma representação dos documentos mais semântica e abrangente. Conclui-se com o estudo que a utilização das ontologias como vocabulários controlados em sistemas de indexação automática pode oferecer

[1] Possui graduação em Engenharia de Computação pela Fundação Educacional de Votuporanga (2010). Licenciado em Computação pelo Centro Universitário Claretiano de Batatais (2013). Mestrado em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (2016). Atualmente é Professor do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Campus Votuporanga na área de Programação e Banco de Dados. Possui experiência com programação para Sistemas Web, utilizando PHP e HTML para desenvolvimento de soluções para os mais diversos propósitos. Trabalha com software livre desde 2013, atuando em um projeto que incentiva o uso de software livre na Universidade. Pesquisador na Área de Ciência da Informação, com ênfase em Indexação Automática e Recuperação de Informação.

[2] Possui graduação em Processamento de Dados pela antiga Fundação Educacional de Bauru (1985). Mestre em Informática pela Universidade Federal da Paraíba (1997). Doutor em Ciências da Comunicação (Ciência da Informação) pela Universidade de São Paulo (2003). Pós-doutorado pela Universidade Federal da Paraíba (2013). Atualmente é professor do Departamento de Ciência da Informação da Universidade Estadual Paulista 'Julio Mesquita Filho' (UNESP) - Campus de Marília. Atua na Ciência da Informação, principalmente nas áreas de Indexação Automática e Recuperação de Informação. Bolsista Produtividade em Pesquisa CNPq - Nível 2.

resultados promissores, permitindo a descoberta automática de termos e a resolução de alguns problemas ligados à linguagem que permeia todo o processo de recuperação de informação.

Palavras-chave: Indexação Automática. Linguagens Documentárias. Ontologias. Recuperação da Informação.

Abstract: *The evolution of technology and access through the internet has changed the way society deals with information and documents. The possibility of storing and sharing the registered knowledge in digital format expands the horizons of science and research. The process of retrieving information studies the relationship between the document collection and thematic representation of the records that compose it and users and their informational needs, which translate into elaborate expressions using a language. An information retrieval system is therefore a mediating element between a documentary collection and its requesters. One of the aspects that directly interferes in their efficiency is how documents are represented. Therefore, researches on automatic indexing take importance, particularly, in an environment of large production and dissemination of documents, as it's the case of the Web. The use of controlled vocabularies as terminology standardization elements is a feature used to improve the results of the indexing process. This study aims to propose and develop a method for using ontologies in the automatic indexing process of textual documents, making use of logical and conceptual structure of domain ontologies and implementing a method that enables automatic indexing systems, an execution of automatic inferences, favoring a semantic and comprehensive documents representation. The study conclusion is that the use of ontologies as controlled vocabularies in automatic indexing systems can offer promising results, allowing the automatic discovery of terms and the resolution of some language related problems that permeates the whole process of information retrieval.*

Keywords: *Automatic Indexing. Controlled Vocabulary. Ontologies. Information Retrieval.*

1. INTRODUÇÃO

O consumo de informações, motivado pelas mais distintas razões é uma realidade habitual para muitos seres humanos. Buscamos conhecimento para satisfazer uma curiosidade, exercer as atribuições profissionais ou simplesmente expandir os próprios conhecimentos e tornarmo-nos mais capacitados para a vida.

A importância do registro do conhecimento produzido pela humanidade é preocupação desde tempos imemoriáveis. Sendo que a representação dos conhecimentos em algum suporte origina documentos, que podem ser armazenados, preservados e disponibilizados. Nesse sentido, as bibliotecas são instituições de grande importância, pois são responsáveis por armazenar, preservar e disseminar estes registros, promovendo assim a multiplicação dos conhecimentos da humanidade.

A difusão da tecnologia e do acesso por meio da internet mudou o modo como a sociedade lida com documentos e informações, pois uma parte significativa destes registros está em formato digital e disponível na Web. Nesse contexto, a Ciência da Informação adquire um importante papel, pois estuda as particularidades do fluxo informacional.

O processo de recuperação de informações pode ser conceituado forma sucinta pela junção de um acervo de documentos, um usuário com necessidades informacionais e um sistema responsável por permitir a seleção de documentos relevantes para este usuário de forma rápida e precisa, satisfazendo assim de forma total ou parcial suas necessidades.

A indexação é um dos processos responsáveis por representar os assuntos de um documento. Por meio de termos o conteúdo temático do documento é representado, sendo que estes servirão como pontos de acesso, mediante os quais um registro será localizado e recuperado no momento da busca.

A linguagem tem um papel fundamental, pois permite que o homem expresse seus pensamentos, ideias e emoções. No entanto, em determinadas situações, ela pode ser incompleta, ambígua e até limitada, o que pode dificultar a compatibilização entre a linguagem utilizada pelo autor e os usuários do sistema de informação. (NARUKAWA, 2011, p.22-23).

A capacidade em lidar com a linguagem de um documento é um dos fatores que determina o grau de êxito de um sistema, sendo que, outros fatores importantes são o método utilizado e o agente responsável pela execução do processo.

A indexação pode ser realizada de forma manual, por humanos, ou de forma automática, por computadores. A primeira abordagem conta com a capacidade de compreensão e cognição humana, que é bastante eficiente ao indicar o assunto de um determinado texto após sua leitura.

No entanto ela sofre de problemas como subjetividade, custo, tempo para realização ou falta de domínio do idioma por parte do indexador (BORGES; MACULAN; LIMA, 2008, p.183). Há ainda a inconsistência entre diferentes indexadores ou para o mesmo indexador em diferentes momentos (BORKO, 1977, apud GUEDES, 1994) resultando em uma redução na qualidade da indexação e consequentemente nos resultados do sistema de recuperação.

Uma solução utilizada para promover uma maior uniformidade no processo de indexação é o uso de vocabulários controlados. Solução esta, que influencia positivamente nos processos de indexação e recuperação da informação, pois amplia o acesso às ideias do autor e serve como elo entre a informação criada e as necessidades dos indivíduos.

A representação da informação realizada através da adoção de terminologias padronizadas resultou em avanços na qualidade da indexação. No entanto, o volume de documentos produzido pela sociedade ainda é um obstáculo para esta forma de indexação.

Visando superar esta dificuldade surgiram as propostas de indexação automática, que exploram a possibilidade de utilizar algoritmos e computadores para realizar a indexação. Com esta abordagem é possível minimizar a subjetividade inerente ao ser humano e indexar grandes volumes de documentos em um curto espaço de tempo.

Várias pesquisas buscam aprimorar a indexação automática utilizando métodos estatísticos a partir da frequência e posição das palavras no documento. No entanto, uma das dificuldades da indexação automática é que vários métodos utilizam como insumo inicial termos extraídos do próprio documento, tornando-os assim suscetíveis às limitações da linguagem.

Uma possível solução é a utilização de vocabulários controlados, tais como os cabeçalhos de assuntos ou tesauros, extensamente utilizados na indexação manual, mas complexos de se aplicar aos sistemas automáticos devido a sua construção inadequada ao uso em sistemas computacionais.

Narukawa (2011) ressalta isso em sua pesquisa sobre o uso de tesauros no sistema de indexação automática SISA, concluindo que "o uso do tesauro oferece benefícios, mas que as falhas, tornam seu uso ainda pouco confiável". A autora aponta como possível solução o uso de ontologias, que, por meio de suas relações podem atender as demandas por sistemas capazes de oferecer tratamento semântico das informações.

A recuperação de informação baseada em ontologia (*ontology-based information retrieval*) tira proveito dos benefícios da utilização destes aparatos capazes de representar o conhecimento e as relações entre os conceitos seguindo uma estrutura formal e processável por computadores.

O objetivo deste trabalho é apresentar um método que utiliza ontologias como vocabulários controlados para expansão automática de termos de indexação. A proposta busca oferecer uma solução auxiliar aos métodos de indexação automática de documentos textuais.

O processo de indexação automática segue basicamente as mesmas etapas da indexação manual, agregando etapas posteriores à extração de termos do documento, como processamentos sintáticos, morfológicos e semânticos (GIL LEIVA, 2008). Na sequência, é feita a etapa de tradução, onde cada candidato a descritor é submetido à comparação utilizando um vocabulário controlado. É neste ponto que o método proposto pode ser utilizado, pois o mesmo apresenta uma abordagem que permite ao sistema expandir as buscas efetuadas no vocabulário controlado a fim de agregar termos relacionados, equivalentes, sinônimos ou ainda em outros idiomas ao conjunto de termos de indexação inicialmente extraídos do documento.

Os resultados obtidos a partir de testes feitos com o protótipo desenvolvido permitem afirmar que as propostas de expansão de termos são tecnicamente factíveis no contexto apresentado, proporcionando assim uma abordagem que pode aprimorar os processos de indexação automática.

2. INDEXAÇÃO AUTOMÁTICA

A indexação é um procedimento que tem por objetivo a elaboração de índices de assunto dos documentos, sendo parte integrante do processo conhecido como análise documentária. A representação dos documentos é uma das etapas do processo de recuperação de informações, sendo a indexação uma das tarefas da representação dos documentos cujo objetivo é indicar do que se trata um determinado documento, como afirma Lancaster (2004, p.6):

> […] a indexação de assuntos e a redação de resumos são atividades intimamente relacionadas, pois ambas implicam a preparação de uma representação do conteúdo temático dos documentos. […] O principal objetivo do resumo é indicar do que se trata o documento ou sintetizar seu conteúdo. Um grupo de termos de indexação serve ao mesmo propósito.

O autor ainda destaca que os termos atribuídos ao índice durante a etapa de indexação servem como pontos de acesso para que o sistema possa localizar no acervo os documentos de interesse dos usuários. Logo, deduz-se que o sucesso desta etapa influencia diretamente os resultados de um sistema de recuperação de informações.

Quanto ao processo de indexação, Lancaster (2004, p. 8-9) explica que há basicamente duas etapas:

- **Análise conceitual do assunto**: requer a compreensão do conteúdo quanto ao tema e tipo de texto para possibilitar a identificação de todos os conceitos que representam o documento visando uma posterior seleção das informações, segundo os parâmetros de exaustividade e especificidade exigidos pelo sistema de indexação.
- **Tradução dos conceitos em termos padronizados:** essa etapa requer a organização das descrições padronizadas de acordo com os termos previamente estabelecidos pelo vocabulário de indexação. É importante salientar que a compatibilidade entre a terminologia de indexação e a de recuperação é fundamental para que os resultados de uma busca sejam satisfatórios.

A análise conceitual do assunto é uma etapa complexa, pois envolve fatores de difícil controle, como a escolha dos termos que representarão os assuntos do documento. Fujita (1989, p.120) afirma que "a escolha de um termo fora de contexto pode prejudicar a recuperação de um documento ao invés de maximizar seu potencial de recuperação no acervo", por isto este processo deve ser norteado por um conjunto de políticas de indexação que levem em consideração fatores como exaustividade, especificidade, coerência, consistência e o público alvo do sistema de recuperação de informações.

Já a etapa de tradução, segundo Lancaster (2004, p.18), pode ser executada de duas formas distintas, dependendo do tipo da indexação. Na indexação por extração, palavras ou expressões do próprio documento são utilizadas para representar o seu conteúdo temático. Já na indexação por atribuição são utilizados termos provenientes de alguma fonte que não é o próprio documento, sendo frequentemente utilizados os vocabulários controlados.

No decorrer deste trabalho adota-se a definição de Lancaster (2004, p. 19) para vocabulários controlados, sendo estes "uma lista de termos autorizados" utilizados pelos indexadores para atribuir a um documento termos que estiverem presentes nesta lista adotada pela instituição. No entanto, o autor afirma que um vocabulário controlado é mais que uma mera lista, pois ele inclui, em geral, uma forma de estrutura semântica, destinada especialmente a controlar sinônimos, diferenciar homógrafos ou ainda reunir termos cujos significados apresentem uma relação mais estreita entre si.

A proposta do uso de vocabulários controlados permite que os indexadores possam se utilizar deste artefato na indexação, minimizando assim possíveis falhas inerentes a subjetividade da interpretação humana.

Dentre os tipos de vocabulários controlados existentes Lancaster (2004, p.19) afirma que

os principais são “esquemas de classificação bibliográfica (como a Classificação Decimal de Dewey), listas de cabeçalhos de assuntos e tesauros”, além destes, existem também as ontologias, que a priori não são vocabulários controlados, no entanto, podem ser utilizadas para tal finalidade.

O uso de ontologias como instrumento de organização da informação vem sendo estudado há algum tempo, principalmente devido a sua característica formal, sendo possível representar os conhecimentos, tal como um tesauro, diferindo deste no sentido de que as ontologias podem ser elaboradas especificamente para uso em sistemas computacionais, tornando-as úteis aos processos de indexação automática.

Após esta breve conceituação sobre o processo de indexação, parte-se para a abordagem da indexação automática, processo realizado por computadores e algoritmos programados para efetuar o processo seleção dos termos mais significativos de um documento, indicando assim seu assunto para que seja possível sua posterior recuperação.

A grande diferença da indexação automática para a indexação realizada por humanos está na objetividade e uniformidade. Um computador é capaz de tomar as mesmas decisões e adotar as mesmas políticas para todos os documentos, independentemente de fatores externos como volume de documentos e conteúdo dos mesmos.

Ferneda (2013) destaca que embora a prática de indexação possa ser regulada por princípios e políticas institucionais, a indexação manual é dependente de critérios subjetivos e pessoais, características inerentes a um indexador humano, mas que afetam negativamente os resultados da indexação por serem difíceis de se controlar. O autor ainda justifica a necessidade de pesquisas e estudos da indexação automática com base na explosão informacional, na quantidade de documentos publicados e do advento da internet, meio extremamente propício a difusão de informações e documentos.

No entanto é preciso considerar que a indexação automática possui falhas, como a dificuldade em lidar com sinônimos ou variações ortográficas, o contexto e a descoberta de assuntos relacionados aos termos candidatos, justificando assim a necessidade de pesquisas e aprimoramentos nos métodos utilizados, como a proposta deste trabalho, que utiliza ontologias como um recurso de auxílio ao processo de indexação automática.

Lancaster (2004, p.18) afirma que há essencialmente dois tipos de indexação: a indexação por extração e a indexação por atribuição. Na primeira os termos são extraídos do próprio documento, o indexador utilizando suas técnicas e seguindo as políticas de indexação deve selecionar termos apenas dentro do documento em questão, selecionando os mais representativos dentro do contexto do documento. Na indexação por atribuição utiliza-se um

elemento externo ao documento, como um conjunto de termos previamente definidos e normalizados, um vocabulário controlado, onde o indexador deve escolher termos que melhor representam o documento a ser indexado (FERNEDA, 2013, p. 84).

De forma semelhante a indexação manual, Lancaster (2004, p.285) explica que a indexação automática também pode ser feita por extração ou atribuição automática. A indexação por extração automática é geralmente implementada utilizando métodos matemáticos de frequência de palavras encontradas no texto, atribuindo ao mesmo os descritores potencialmente mais significativos.

Na indexação por atribuição automática existe a necessidade de se fazer uma indexação por extração como uma primeira etapa, objetivando obter um conjunto de candidatos a descritores retirados do próprio documento. Em seguida cada um destes descritores deve ser submetido a um vocabulário controlado, para assim obter a forma normalizada, ou ainda outros termos relacionados, possibilitando a escolha de descritores padronizados, que podem resultar em melhores resultados na recuperação da informação.

O processo de indexação por atribuição automática possui uma maior complexidade envolvida, pois escolher termos que representem o assunto de um determinado documento requer uma análise do contexto, sendo esta uma operação subjetiva, logo, mais difícil de ser automatizada.

A utilização de ontologias no processo de indexação por atribuição automática se justifica pela necessidade de um elemento externo que contenha um conjunto de termos para servir aos propósitos da etapa de tradução da indexação. As ontologias destacam-se para esta finalidade pois são artefatos que tratam da organização de objetos, suas propriedades e seus relacionamentos em um determinado domínio de conhecimento (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999).

As ontologias abrem novas perspectivas para o desenvolvimento de sistemas de indexação automática pois oferecem uma estrutura conceitual e terminológica restrita a um domínio e representadas em linguagens processáveis por máquinas, tal como XML, o que permite sua utilização nos mais variados processos computacionais (FERNEDA, 2013 p. 90).

Ainda segundo o autor, a utilização de uma base ontológica possibilita uma abordagem mais rica para a indexação, pois oferece algum nível de análise semântica. Após extrair os termos de um documento é possível mapeá-los em uma ontologia, padronizando o vocabulário e contextualizando de forma mais eficiente os termos no domínio da ontologia, solucionando assim possíveis ambiguidades.

3. ONTOLOGIAS

O termo ontologia possui definições distintas dependendo da área de conhecimento ao qual é citado. No âmbito da ciência da computação uma definição clássica é da de Gruber (1995), para o qual uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada. Onde, "formal" diz respeito a "ser legível por computador"; "explícita", indica que os elementos estão claramente definidos; "conceitualização" refere-se a um modelo abstrato de um fenômeno; e compartilhada significa que os conceitos presentes representam um conhecimento consensual, aceito por um grupo de pessoas.

Uma ontologia pode ser considerada como um vocabulário de representação, geralmente especializado em um domínio ou assunto, qualificado por conceituações de tipos de objetos e suas relações no mundo. Em outras palavras, é um corpo de conhecimento que descreve algum domínio, utilizando um vocabulário de representação (CHANDRASEKARAN; JOSEPHSON; BENJAMIN, 1999).

Jasper e Uschold (1999, p.11-2) destacam um ponto crucial do uso de ontologias, a possibilidade de criar restrições, explicitar as relações entre os termos, agregar rótulos em múltiplos idiomas conceituando assim cada termo com expressividade suficiente para que este sejam compreensíveis tanto por humanos quanto por máquinas com o mesmo nível de compreensão.

Segundo Soergel (1999) e Vickery (1997), o termo ontologia começou a ser utilizado na literatura da Ciência da Informação no final da década de 1990, mas principalmente por pesquisadores da área de Organização do Conhecimento. Nesta época, os instrumentos e métodos de classificação despertavam maior interesse da Computação, devido principalmente a necessidade de desenvolvimento de instrumentos de organização da informação na Web.

Entre os instrumentos de representação tradicionalmente utilizados na área de Ciência da Informação, os tesauros são os mais semelhantes às ontologias. Como destacam Sales e Café (2008), ambos os instrumentos são constituídos por linguagens de estruturas combinatórias, de caráter especializado, representando termos e conceitos organizados a partir de tipos de relacionamentos, no entanto, estudos têm constatado que, apesar de possuírem características em comum, tais instrumentos constituem-se como diferentes modelos de representação do conhecimento.

Pickler (2014, p.61) destaca que um fator determinante na distinção das ontologias para outros modelos de representação da informação é que as relações expressas nas ontologias são formalmente descritas, permitindo assim a realização de inferências automáticas. Desta forma o

potencial de representação e expressividade proporcionado pelo uso de ontologias as colocam como um valioso instrumento de representação da informação para a Ciência da Informação.

Este estudo busca estreitar as relações entre as ciências da informação e da computação, pois se presta a utilizar ferramentas e softwares para verificar os resultados de uma indexação automática utilizando ontologias como ferramentas de apoio a decisão em situações dúbias durante a indexação automática de documentos textuais.

A ontologia utilizada neste trabalho é escrita em linguagem OWL, que é recomendada pelo consórcio W3C como a principal linguagem para a construção de ontologias. Ela tem como objetivo principal atender às necessidades de aplicação da Web Semântica e ser efetivamente utilizada por aplicações que necessitem processar o conteúdo de informações, e não somente apresentar a visualização destas informações. (SANTARÉM SEGUNDO, 2010, p.127).

A documentação oficial disponibilizada pelo W3C afirma que a OWL é uma linguagem de web semântica projetada para representar o conhecimento em toda sua riqueza e complexidade, além de grupos de coisas e as relações entre as coisas. Ela é baseada em lógica computacional onde o conhecimento é expresso de forma que o computador possa explorá-lo e verificá-lo sistematicamente, além de ter acesso a informações explícitas ou implícitas por meio de inferências. A OWL é parte da "pilha" de tecnologias propostas pelo W3C, que inclui ainda RDF, RDFS, SPARQL, entre outras.

A linguagem de consulta SPARQL é definida Prud'Hommeaux (2008, tradução nossa) como uma linguagem para expressar consultas em diversas fontes de dados, desde que os dados estejam armazenados nativamente como RDF. Segundo Luz (2013) o termo SPARQL, cuja pronúncia é "sparkle" é um acrônimo recursivo para SPARQL *Protocol and RDF Query Language* corroborando com a definição de Prud'Hommeaux sobre a proposta de ser um mecanismo de consulta a dados estruturados no formato *Resource Description Framework* (RDF).

Filho e Lóscio (2009) colocam que a linguagem SPARQL é a linguagem recomendada pelo W3C para a recuperação de informações em documentos no formato RDF/RDFS, tal como a linguagem SQL é usada para recuperar registros em bases de dados estruturados.

Um processador SPARQL é um software capaz de processar consultas em dados locais ou remotos DuCharme (2013, tradução nossa). Para os testes realizados foi utilizado o Jena framework, um programa baseado em Java capaz de realizar consultas em linguagem SPARQL em ontologias no formato OWL.

A proposta deste trabalho utiliza apenas alguns recursos da linguagem SPARQL e as ontologias OWL como vocabulários controlados, sendo que as ontologias possuem um amplo

potencial de utilização, mas que foge ao escopo do trabalho.

4. MÉTODO DE EXPANSÃO DE TERMOS DE INDEXAÇÃO BASEADO EM ONTOLOGIA

Neste capítulo descreve-se o conceito e o funcionamento do método de expansão de termos de indexação baseado em ontologias, apresentando na sequência os materiais utilizados no desenvolvimento do protótipo construído para simular o funcionamento do método.

Para compreender o método proposto é necessário ter em mente o objetivo e o processo de indexação, sendo que Gil Leiva (2008, p.63, tradução nossa) afirma que "o objetivo da indexação de documentos é permitir o armazenamento das informações para atender as necessidades informacionais dos indivíduos", apresentando uma visão do processo de indexação composta basicamente por duas etapas:

1 – Extração e Mapeamento dos conceitos

2 – Tradução.

A definição do processo de indexação composta por duas etapas é descrita também por UNISIST (1981, p.85), Chaumier (1988, p.23), Fidel (1994, p.573) e Lancaster (2004, p.8-9), demonstrando assim que esta é aceita por alguns autores. No entanto, há divergências entre os autores, sendo que em alguns casos o processo chega a ter até cinco etapas. Estas divergências residem principalmente na subdivisão das etapas, sendo que essencialmente o processo é composto pelas etapas de extração e mapeamento de conceitos seguido de tradução.

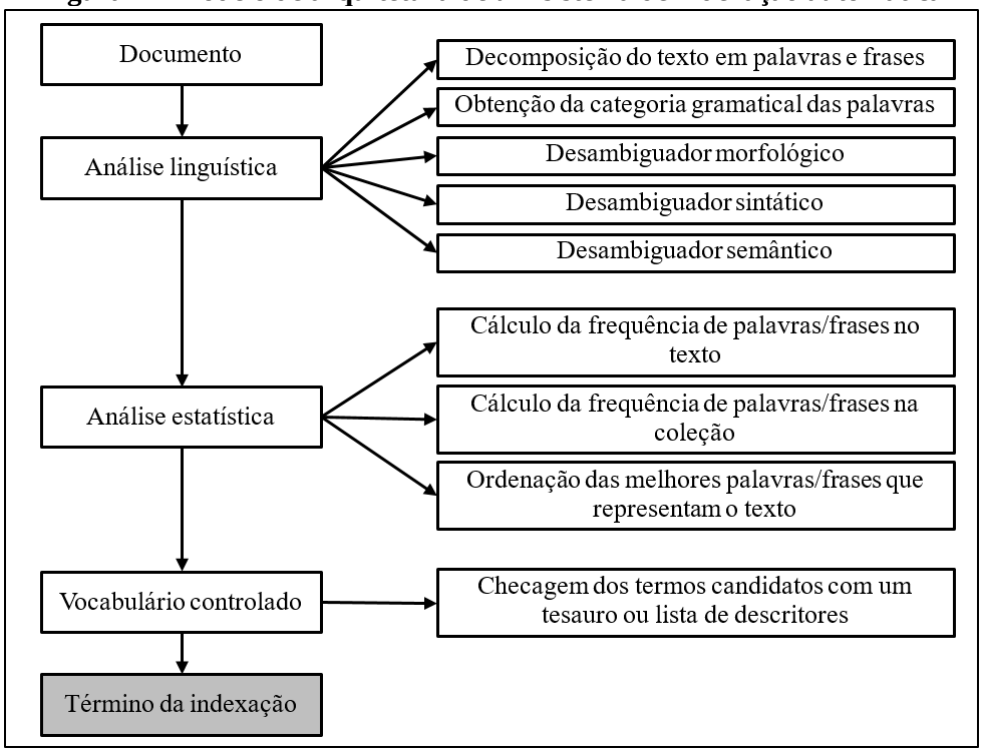
Em relação ao processo de indexação automática, Gil Leiva (2008, p.368) afirma que também não há um consenso, devido principalmente a complexidade do assunto, no entanto, o autor apresenta um modelo de arquitetura de sistema de indexação automática, que pode ser observado na Figura 1, onde algumas etapas comuns são apresentadas baseadas em estudos de propostas levantadas pelo autor entre os anos 1950 e 2008.

A partir do modelo de arquitetura proposto por Gil Leiva é possível observar três momentos distintos, a análise linguística, responsável pelo tratamento primário das palavras ou frases do texto, a análise estatística, onde por meio de algoritmos e/ou regras pré-definidas são escolhidos os candidatos a descritores do documento e, por fim, a checagem destes utilizando um vocabulário controlado, etapa ao qual o autor cita o uso de tesauros ou listas de descritores.

Em sua tese de doutorado Gil Leiva propôs um sistema de indexação automática, denominado SISA, onde um tesauro é utilizado como ferramenta de controle terminológico na etapa de cotejamento dos termos candidatos. Narukawa (2011) faz uma análise comparativa dos resultados da utilização do SISA para indexação automática e da indexação manual de um

corpo de documentos, concluindo que a utilização do tesauro como vocabulário controlado torna a solução susceptível de falhas, indicando ainda como uma possível solução o uso de ontologias em substituição ao tesauro.

Figura 1 – Modelo de arquitetura de um sistema de indexação automática



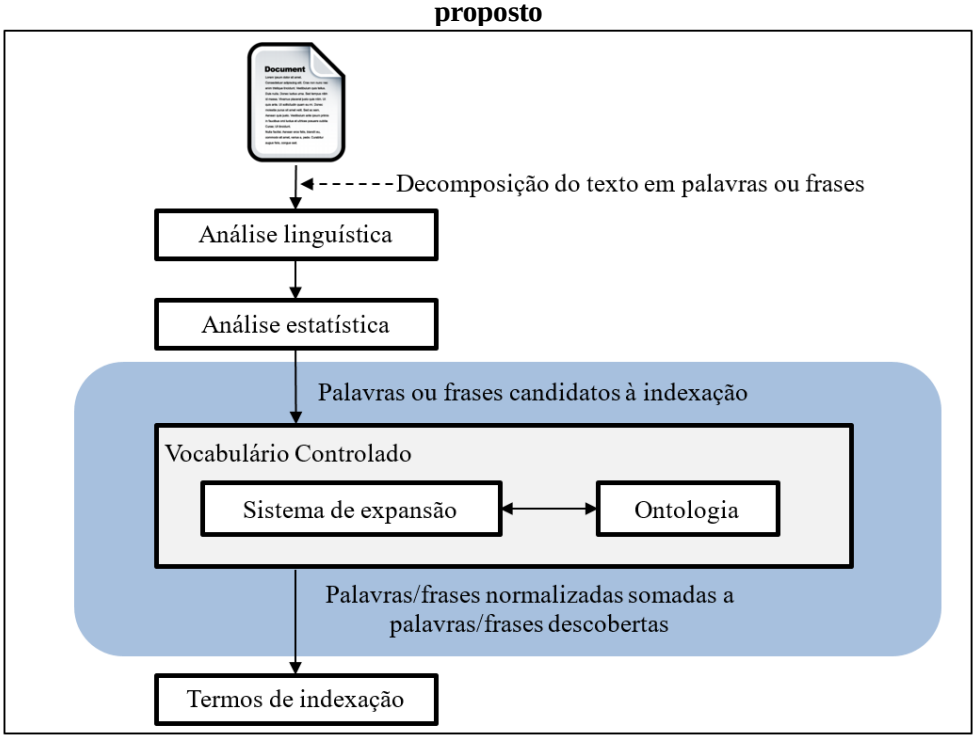
Fonte: adaptado de Gil Leiva (2008, p.367)

Desta forma o método proposto busca atacar essa problemática, pois, pode auxiliar no processo de indexação automática junto a etapa de cotejamento de termos no vocabulário controlado. Destaca-se que a proposta do trabalho não é desenvolver um sistema de indexação automática, mas um artefato auxiliar ao processo.

Considerando a arquitetura de um sistema de indexação automática proposta por Gil Leiva e apresentada na Figura 1, o método proposto busca aperfeiçoar a terceira etapa, onde uma checagem dos termos candidatos é feita utilizando um tesauro ou uma lista de descritores, sendo que a proposta desta pesquisa utiliza uma palavra ou frase como entrada, e, por meio do uso de uma ontologia de domínio resulta em uma ou mais palavras ou frases em sua saída.

Uma representação de um processo de indexação automática genérico, baseado no modelo de Gil Leiva (2008) é mostrada na Figura 2, sendo que o quadro em destaque representa a aplicação do método proposto no processo de indexação automática.

Figura 2 – Modelo de arquitetura do processo de indexação automática utilizando o método proposto



Fonte: desenvolvido pelo autor

O funcionamento do método destacado no quadro da Figura 2 consiste em uma aplicação capaz de receber um termo candidato a descrito, fazer sua expansão utilizando uma ontologia de domínio e retornar o termo original juntamente com os novos descobertos.

A ontologia de domínio utilizada neste trabalho foi a PedTerm, uma ontologia que representa o domínio de conhecimento da pediatria apresentando informações relacionadas à saúde e ao desenvolvimento infantil desde o pré-natal até os 21 anos de idade.

Esta ontologia, criada originalmente no idioma Inglês com suas classes e propriedades nomeadas apenas usando o atributo ID, não foram utilizados rótulos (propriedade *label*) OWL para descrever suas entidades. Isto configura um problema para a abordagem proposta, pois os métodos apresentados utilizam os rótulos para comparar os termos do documento com o vocabulário controlado.

Desta forma foi necessário fazer alguns ajustes na ontologia original, inserindo os *labels* em algumas classes. Nesta etapa houve-se o cuidado de criar rótulos na língua original da ontologia (Inglês) e também em português e espanhol, para que fosse possível testar as propostas de indexação multilíngue.

O método proposto utiliza ontologias como ferramentas de controle terminológico em uma indexação por atribuição. No entanto, antes de efetuar a atribuição de termos a um determinado documento é necessária uma etapa de extração, sendo que primeiramente são

extraídos do documento os candidatos a descritores, para depois submeter estes ao sistema que é responsável por indicar outros termos relacionados ao descritor candidato.

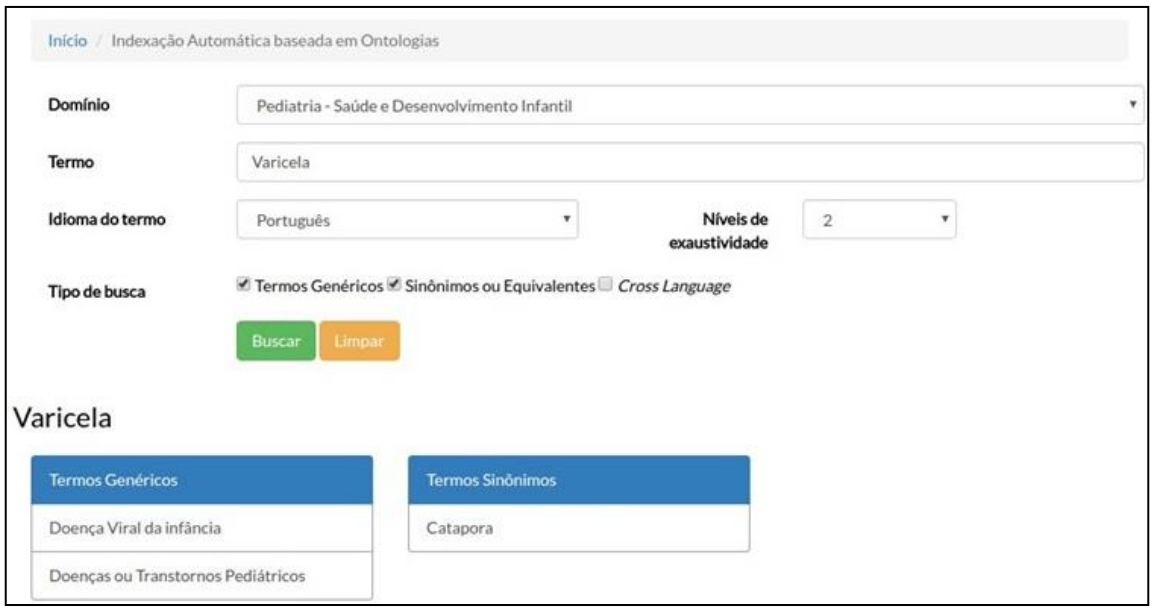
A utilização de ontologias no processo de indexação automática implica em determinar o assunto dos documentos (*corpus*), desta forma, ao definir que um documento está relacionado a uma ontologia de domínio define-se implicitamente que o documento trata do assunto desta ontologia. Para a proposta deste estudo assume-se que os termos em questão e a ontologia são de um mesmo domínio, sendo que nenhuma abordagem de descoberta automática de assunto ou utilização de múltiplas ontologias é explorada.

Nos próximos parágrafos demonstra-se a utilização do protótipo para expandir termos hipoteticamente extraídos de documentos textuais. Primeiramente são explicadas as propostas e resultados esperados de cada teste, em seguida são apresentados os parâmetros de entrada e os resultados obtidos após a execução do teste.

O funcionamento do protótipo implementado consiste em primeiramente escolher a ontologia de domínio por meio do campo "Domínio", o primeiro campo que pode ser visto na Figura 3. Depois deve-se informar no campo "Termo" uma palavra que seria um candidato a descritor hipotético, em seguida escolhe-se o idioma do termo informado, para que o sistema saiba em quais rótulos fazer a busca, e por fim deve-se selecionar o idioma do termo informado.

O nível de exaustividade se refere a quantos níveis hierárquicos o sistema deverá buscar, caso seja escolhido um nível, apenas uma classe hierarquicamente superior será analisada, caso seja escolhido dois níveis, duas classes hierarquicamente superiores serão analisadas, e assim sucessivamente.

Figura 3 – Interface do sistema exemplificando um teste com o termo "Varicela"



Fonte: desenvolvido pelo autor

A última etapa é escolher no campo “tipo de busca”, que se refere ao tipo de expansão de termos que será executada pelo sistema. A opção “Termos Genéricos” faz com que o sistema busque por termos hierarquicamente genéricos a partir do termo informado, obedecendo ao nível de exaustividade previamente escolhido. A opção “Sinônimos ou Equivalentes” busca em classes equivalentes e em classes com dois ou mais rótulos representando um mesmo conceito, resultando assim em sinônimos ou em termos equivalentes em relação ao informado.

A última opção “*Cross Language*” informa ao sistema para que deixe de filtrar os resultados apenas no idioma do termo informado, apresentando como resultado os termos de todos os idiomas contemplados pela ontologia correspondentes às relações de hierarquia ou equivalência encontradas.

O método de expansão de termos baseado em ontologia desta proposta baseia-se em três abordagens, a **atribuição de conceitos hierarquicamente genéricos**, a **descoberta de termos sinônimos** e a **indexação multilíngue**. A atribuição de conceitos funciona da seguinte forma: cada termo extraído do texto é submetido por meio de uma consulta SPARQL a uma ontologia, buscando coincidência entre este e o rótulo de alguma classe. Quando uma coincidência ocorre, a classe é adotada como representante do conceito de entrada servindo de ponto de partida para a realização de inferências que permitem traçar relacionamentos com outras classes.

A Figura 4 apresenta-se um exemplo de funcionamento do método abordando a expansão por meio da atribuição de conceitos hierarquicamente genéricos e a indexação multilíngue. O termo inicial “Hipotireoidismo” é submetido ao sistema, que localiza a partir dele seus correspondentes no idioma inglês e espanhol, na sequência faz-se uma busca por classes hierarquicamente superiores, onde é encontrada a classe “*Endocrine System Disorder*” que possui como rótulo “Transtorno do Sistema Endócrino” além dos rótulos em inglês e espanhol. A mesma lógica é repetida em um segundo nível de exaustividade.

Este exemplo hipotético poderia ser utilizado para indexar diferentes documentos nos idiomas inglês, espanhol e português, como representam as figuras à esquerda da Figura 4, no entanto nada impede que uma indexação multilíngue seja executada em um único documento.

Um teste feito utilizando esta lógica utiliza como insumo inicial o termo “Tétano”, uma doença infeciosa causada por bactérias, foi possível observar que o termo pertence a ontologia, representado pela classe “*Tetanus*”, que por sua vez é uma subclasse de “*Bacterial_Disease*”, e esta também é uma subclasse de “*Infectious_Disease*”. Desta forma, utilizando o protótipo com uma exaustividade de dois níveis foi possível expandir o termo original “Tétano” para “Tétano”, “Doença bacteriana” e “Doença infecciosa”.

Figura 4 – Interface do sistema exemplificando um teste com o termo "Varicela"

Fonte: Adaptado de PICKLER, 2014, p.98

Outro teste, que explora a proposta de descoberta de termos sinônimos ou equivalentes, utiliza dois recursos da linguagem OWL, o primeiro é a liberdade de declarar um ou mais rótulos para uma mesma classe. O segundo é a relação de equivalência entre classes, que possibilita explicitar que duas classes representam conceitos equivalentes. Neste teste o termo inicial é "Varicela", após submetido ao protótipo a classe "*Chicken_Pox*" é localizada contendo um rótulo "Varicela" e também um segundo rótulo "Catapora", sendo ambos os termos retornados como uma expansão do termo inicial.

Na Figura 3 pode-se observar a interface do sistema com a execução de um teste cujo termo inicial é "Varicela". No resultado, que pode ser visto na parte inferior da imagem, há o termo original "Varicela" juntamente com os termos "Doença Viral da infância" e "Doenças ou Transtornos Pediátricos", que são termos hierarquicamente genéricos e também o termo sinônimo "Catapora". Os termos genéricos e sinônimos são mostrados em painéis diferentes apenas para facilitar a visualização dos resultados dos testes.

Um novo teste busca explorar a relação de equivalência entre classes, partindo do termo "Espinha" o protótipo localiza a classe "*Pimple*" que possui uma classe equivalente "Acne", desta forma o termo inicial "Espinha" é expandido para "Espinha" e "Acne".

É importante frisar que consideramos a relação associativa de termos "sinônimos" e "equivalentes" baseado nas definições provenientes da área computacional, especificamente da linguagem OWL, que define uma equivalência entre classes como "uma construção que pode

ser usada para criar classes sinônimas para ligar duas classes em uma mesma ontologia ou em ontologias diferentes" (CARVALHO; LIMA, 2005).

Ambos os testes exploram a descoberta de termos sinônimos ou equivalentes permitem demonstrar uma possível solução aplicável a desambiguação automática de termos em sistemas de indexação automática, sendo esta uma dificuldade comum em processos de indexação automática.

Por fim, a indexação multilíngue, que também explora a possibilidade de criar um ou mais rótulos para uma mesma classe, e, utilizando o parâmetro *xml:lang* é possível que uma mesma classe seja descrita em vários idiomas, com isso uma ontologia pode ser utilizada para indexação de um corpus contendo documentos em diferentes idiomas.

No protótipo desenvolvido há um campo "Idioma do termo" que serve para definir qual o idioma do termo informado como entrada, sendo que a partir dele é possível filtrar os resultados da expansão de termos hierarquicamente genéricos ou termos sinônimos ou equivalentes para o idioma de entrada. Ou ainda, deixar de utilizar o filtro de idioma de saída, sendo que nesta abordagem é possível que um termo informado no idioma português seja traduzido automaticamente para todos os idiomas suportados pela ontologia.

Testes foram realizados no protótipo e permitiram verificar que a proposta funciona, a partir da entrada de termos em um idioma é possível retornar este expandido em todos os idiomas declarados na ontologia.

Destaca-se que todas as propostas de expansão de termos desenvolvidas no protótipo podem ser utilizadas em conjunto, sendo que foram feitos testes que provaram ser possível fazer uma expansão de termos hierarquicamente genéricos e buscando sinônimos ou ainda descobrir sinônimos ou equivalentes de um termo e vários idiomas além de outras combinações. Isto demostra a flexibilidade do método desenvolvido.

5. CONSIDERAÇÕES FINAIS

As pesquisas em sistemas de indexação automática têm por objetivo tornar o processo de representação de documentos mais ágil, pois, considerando a quantidade de documentos disponíveis é possível compreender a grande demanda por mecanismos que permitam a automatização deste processo.

O processo de indexação é uma tarefa que lida com elementos de natureza linguística, sofrendo influência de fatores subjetivos e de difícil formalização. Neste contexto, o uso de vocabulários controlados possibilita a obtenção de melhores resultados.

Assim como na indexação manual, a indexação automática também apresenta benefícios com a utilização de vocabulários controlados. No entanto, é necessário buscar opções de vocabulários que sejam adequados ao uso por sistemas computacionais, tal como as ontologias.

A utilização de ontologias vem de encontro a necessidade de vocabulários controlados processáveis por máquinas, tal como afirma Narukawa (2011), a adoção de ontologias em substituição de outros vocabulários, como tesauros, em sistemas de indexação automática pode prover mais confiabilidade ao sistema, pois estas proveem um tratamento semântico das informações.

Para que uma ontologia seja utilizada como um vocabulário controlado, é necessário que ela seja construída para esta finalidade, ou, no caso do uso de ontologias previamente elaboradas, há a necessidade de ajustes, lembrando que, para a proposta apresentada por esta pesquisa a declaração dos rótulos dos conceitos (*labels*) com seus respectivos idiomas torna-se obrigatória.

Os resultados obtidos a partir do protótipo desenvolvido no contexto deste trabalho permitem concluir que, o uso de ontologias como ferramenta de controle terminológico em sistemas de indexação automática pode aprimorar o processo de indexação, implicando em melhoras na recuperação de informação como um todo.

Sendo o objetivo desta pesquisa propor e desenvolver um método de utilização de ontologias em uma das etapas do processo de indexação automática de documentos textuais, é possível afirmar, a partir dos resultados obtidos, que a proposta é factível e aplicável, especificamente junto a etapa posterior a extração dos termos candidatos a descritores fornecendo meios para análise e padronização do índice.

Destaca-se a importância do protótipo construído, pois, por meio das pesquisas para seu desenvolvimento foi possível observar a importância da utilização das tecnologias e propostas da web semântica no desenvolvimento de sistemas de informação e comunicação mais inteligentes.

A partir das leituras e comparações com outros trabalhos correlatos foi possível observar que há muitos autores discutindo sobre o uso de ontologias em processos de indexação, mas poucos apresentam propostas concretas ou resultados sólidos da real aplicação dessas propostas.

Alguns trabalhos, de áreas como a computação, possuem um enfoque mais prático, com a proposta de criar sistemas utilizando ontologias em seu desenvolvimento, mas, sem explorar em profundidade temas como a importância do processo de indexação e da definição das políticas de indexação, fatores estes que influenciam o funcionamento dos sistemas de recuperação de informação.

Por fim, este trabalho busca ser uma colaboração para os estudos em indexação automática utilizando ontologias como vocabulários controlados, pois apresenta, além dos resultados obtidos, uma metodologia bastante clara de como toda a implementação foi feita, visando possibilitar que trabalhos futuros se utilizem dos conhecimentos apresentados para desenvolver novas e mais eficientes técnicas que possam cada vez mais melhorar a indexação e, consequentemente, a busca por informações relevantes.

REFERÊNCIAS

BORGES, Graciane Silva Bruzinga; MACULAN, Benildes Coura Moreira dos Santos; LIMA, Gercina Angela Borem de Oliveira. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade**: Estudos, v. 18, n. 2, p.181-193, 2008. Disponível em: < http://goo.gl/FMBsb2 >. Acesso em: 06 abr. 2016.

BORKO, H. Toward a theory of indexing. Information Processing and Management, v.13, p. 355-365, 1977 apud GUEDES, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da Informação**, Brasília, DF, v. 23, n. 3, p.318-326, set./dez. 1994.

CHANDRASEKARAN, B.; JOSEPHSON, J.R.; BENJAMINS, V. R.1999. What are ontologies, and why do we need them ? **IEEE Intelligent Systems**, v.14, n.1, 1999.

CHAUMIER, Jacques. Indexação: conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. 1/2, p. 63-79, 1988. apud GIL LEIVA, I. **Manual de indización**: teoría y práctica. Gijón: Trea, 2008. 429 p.

DUCHARME, Bob. **Learning SPARQL**: Querying and Updating with SPARQL 1.1. 2. ed. Sebastopol (USA): O' Reilly, 2013.

FERNEDA, Edberto. **Modelos computacionais de recuperação de informação**. Rio de Janeiro: Ciência Moderna, 2012.

_____. **Ontologia como recurso de padronização terminológica em um Sistema de Recuperação de Informação**. 2013. 109 f. Relatório (Estágio Pós-Doutoral), Ciência da Informação, Universidade Federal da Paraíba, João Pessoa. 2013.

FIDEL, Raya. User-centered indexing. **Journal of the American Society for Information Science** (1986-1998), v. 45, n. 8, p. 572, 1994.

FILHO, F.W.; LÓSCIO, B.F. Web Semântica: Conceitos e Tecnologias. In: Pedro de Alcantara. (Org.). **II Escola Regional de Computação** - Ceará, Maranhão e Piauí - ERCEMAPI 2009. : SBC, 2009.

FUJITA, Mariângela Spotti Lopes. Avaliação da eficácia de recuperação do sistema de indexação PRECIS. **Ciência da Informação**, [S.l.], v. 18, n. 2, Dez. 1989. ISSN 1518-8353. Disponível em: < http://goo.gl/fxgXZP >. Acesso em: 07 Set. 2015.

GIL LEIVA, I. **Manual de indización**: teoría y práctica. Gijón: Trea, 2008. 429 p.

GRUBER, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal Human-Computer Studies,** v.43, n.5-6, 1995.

KARA, Soner et al. An ontology-based retrieval system using semantic indexing. **Information Systems**, [s.l.], v. 37, n. 4, p.294-305, jun. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.is.2011.09.004.

LANCASTER, F.W. **Indexação e Resumos**: teoria e prática. 2.ed. Brasília, DF: Briquet de Lemos, 2004.

LIMA, J. C.; CARVALHO, C. L. **Ontologias – OWL (Web Ontology Language)**. 2005. Relatório Técnico - Instituto de Informática Universidade Federal de Goiás. Disponível em: < http://goo.gl/8vyctb >. Acesso em: 28 set. 2015.

NARUKAWA, C. Miyuki. **Estudo de vocabulário controlado na indexação automática**: aplicação no processo de indexação do Sistema de Indización Semiautomática (SISA). 2011. 222 f. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2011.

PICKLER, Maria Elisa Valentim. **Diretrizes para utilização de ontologias na indexação automática**. 2014. 101f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.

PRUD'HOMMEAUX, Eric et al. SPARQL query language for RDF. **W3C recommendation**, v. 15, 2008.

SANTARÉM SEGUNDO, J E**. Representação Iterativa**: um modelo para repositórios digitais. 2010. 224f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

SOERGEL, D. The rise of ontologies or the reinvention of classification. **Journal of the American Society for Information Science**. v. 50, n. 12, 1999.

UNISIST. Princípios de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v.10, n.1, p. 83-94, mar. 1981.

VICKERY, Brian C. Ontologies. **Journal of information science**, v. 23, n. 4, p. 277-286, 1997.