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

EDER ANTONIO PANSANI JUNIOR

CONTEXTUALIZAÇÃO E EXPANSÃO DE CONSULTAS EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO:

Um método baseado em ontologias de domínio

MARÍLIA
2021

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

EDER ANTONIO PANSANI JUNIOR

CONTEXTUALIZAÇÃO E EXPANSÃO DE CONSULTAS EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO:

Um método baseado em ontologias de domínio

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências - Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, campus de Marília, como requisito parcial para obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Informação, Tecnologia e Conhecimento.

Linha de pesquisa: Informação e Tecnologia.

Orientador: Prof. Dr. Edberto Ferneda.

MARÍLIA
2021

P196c

Pansani Junior, Eder Antonio
Contextualização e expansão de consultas em sistemas de recuperação de informação : Um método baseado em ontologias de domínio / Eder Antonio Pansani Junior. -- Marília, 2021
162 p.

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

1. Recuperação da informação. 2. Sistemas de recuperação da informação. 3. Query (Sistema de recuperação da informação). 4. Ontologia. 5. ContextOnSearch. 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.

EDER ANTONIO PANSANI JUNIOR

CONTEXTUALIZAÇÃO E EXPANSÃO DE CONSULTAS EM SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO:
Um método baseado em ontologias de domínio

BANCA EXAMINADORA

Prof. Dr. Edberto Ferneda (orientador)
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Prof. Dr. José Eduardo Santarém Segundo
Programa de Pós-Graduação em Ciência da Informação (UNESP)
Departamento de Educação, Informação e Comunicação
Universidade de São Paulo (USP)

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

Profa. Dra. Márcia Cristina dos Reis
Instituto Federal do Paraná, Campus Jacarezinho

Prof. Dr. Guilherme Ataíde Dias
Departamento de Ciência da Informação
Universidade Federal da Paraíba (UFPB)

Marília, 27 de maio de 2021.

Dedico este trabalho à minha esposa Amanda, o amor da minha vida. Minha melhor amiga e companheira, que esteve presente durante as etapas mais difíceis de minha vida me apoiando e incentivando.

Dedico também à minha família, em especial a minha mãe Márcia e a minha sogra Iara, que considero minha segunda mãe.

AGRADECIMENTOS

Agradeço primeiramente à **Deus**, por ter me abençoado com saúde e força para trabalhar e superar as dificuldades e desafios cotidianos.

A minha esposa **Amanda**, pela compreensão e paciência, nos muitos momentos em que meu tempo e atenção foram dedicados aos estudos e a pesquisa. Seu apoio foi fundamental para que eu perseverasse no desenvolvimento do trabalho e tivesse condições de terminá-lo.

A minha filha **Liz**, que, mesmo com tão pouca idade, tem me ensinado diariamente como o amor pode ser terno e revigorante.

Aos meus pais, **Eder** e **Márcia**, que desde cedo nunca mediram esforços para que eu pudesse estudar e me qualificar, proporcionando condições para que isso fosse possível. Eu não tenho como lhes agradecer, mas agradeço à Deus por ter vocês ao meu lado neste momento da vida.

A familiares e amigos, que por meio de seu apoio e amizade tornaram esta longa jornada mais agradável, oferecendo uma palavra, um apoio ou simplesmente a sua presença.

Ao meu orientador, **Prof. Dr, Edberto Ferneda**, pela sua amizade, paciência e confiança em mim depositadas. Agradeço sua disponibilidade em auxiliar no desenvolvimento da tese, incentivando e apoiando meu trabalho. As inúmeras revisões e sugestões dadas ajudaram a lapidar as arestas desta pesquisa. Agradeço ainda pelos almoços no shopping, que mesclavam momentos de orientação e descontração na medida certa. Obrigado por ter me ensinado como conduzir uma orientação de forma saudável, permeada pela amizade e responsabilidade. Você é um ótimo professor e uma pessoa fantástica, que fez uma grande diferença em minha vida.

Aos professores **Dr. José Eduardo Santarém Segundo** e **Dra. Márcia Cristina dos Reis** pelas sugestões e críticas oferecidas durante o exame de qualificação, que contribuíram de forma substancial no direcionamento, amadurecimento e finalização deste estudo.

Aos professores **Dr. Walter Moreira** e **Dr. Guilherme Ataíde Dias**, que aceitaram o convite para integrar a banca avaliadora da defesa, dispondo de seu valioso tempo para fazer uma leitura crítica de meu trabalho com vistas ao seu aprimoramento.

À **Richele Grenge Vignoli** e a **Maria Elisa Valentim Pickler Nicolino** pelo auxílio com a normalização acadêmica do trabalho, principalmente em relação à organização das referências bibliográficas e o ajuste de elementos textuais;

A todos os **docentes do Programa de Pós-Graduação em Ciência da Informação**, pelos conhecimentos transmitidos ao longo destes anos de convivência, que foram esclarecedores e expandiram de forma irreversível minha visão de mundo.

Aos colegas e amigos do **Programa de Pós-Graduação em Ciência da Informação da UNESP – Campus Marília**, que me ajudaram a entender a pós graduação e suas nuances. Todos vocês me proporcionaram, além do aprendizado, companheirismo e troca de experiências, imprescindíveis à consecução desta empreitada.

Ao **Instituto Federal de São Paulo (IFSP)**, que me concedeu afastamento integral durante a elaboração da tese, permitindo uma maior e melhor dedicação às atividades acadêmicas do doutorado e colaborando significativamente com a qualidade final do relatório.

Aos professores e servidores do **Instituto Federal de São Paulo (IFSP) – Campus Votuporanga**, pela colaboração e apoio durante meu afastamento das atividades de ensino, pesquisa e extensão.

Por fim, agradeço a todos que contribuíram de forma direta ou indireta no desenvolvimento desta tese, tornando possível sua consecução.

Muito Obrigado!

“Informação é uma fonte de aprendizagem.
Mas, se não está organizada, processada
e disponível para as pessoas certas
em um formato que ajude a tomar decisões,
é uma carga, não um benefício.”

William Pollard

RESUMO

Um sistema de recuperação de informação é formado por três elementos básicos: as representações dos documentos, a expressão de busca do usuário e alguma forma de comparação entre esses dois elementos. Por um lado, o acervo é constituído em um momento anterior às buscas, e cada documento pode ser representado utilizando técnicas automatizadas. Por outro lado, a necessidade de informação do usuário só é percebida após a sua enunciação por meio de expressão de busca. A elaboração de uma expressão de busca, que represente de forma precisa a necessidade de informação de um usuário pode ser uma tarefa complexa. Nesse sentido, as ontologias no papel de instrumentos de controle de vocabulário podem ser utilizadas no aprimoramento desta tarefa. As ontologias possibilitam, entre outras funções, a contextualização da busca, utilizando sua estrutura terminológica para procurar pelos termos/conceitos que compõem à consulta pode-se determinar o contexto da busca. A partir desse contexto, selecionado pelo usuário e representado por uma ontologia, pode-se expandir a consulta utilizando termos/conceitos relacionados. Desta forma, esta pesquisa tem como objetivo geral a proposição de um método interativo de contextualização da necessidade de informação e expansão de consultas em sistemas de recuperação de informação utilizando a estrutura terminológica das ontologias de domínio. Como objetivos específicos, a tese se propõe a: a) discutir o processo de Recuperação de Informação, evidenciando a relação entre a expressão de busca e os resultados recuperados; b) discutir os conceitos e características das ontologias, explorando sua utilidade nos processos de recuperação de informação, expansão e contextualização das consultas; c) desenvolver um método de contextualização da busca, por meio dos termos que compõem a consulta e expansão da consulta a partir da identificação de conceitos relacionados (genéricos, específicos e equivalentes) à um conceito inicial, utilizando para ambos ontologias de domínio. d) implementar um protótipo de um sistema de recuperação de informação Web para demonstrar a utilização do método proposto em um ambiente controlado; e) analisar os resultados obtidos em relação à relevância com a expressão de busca. Esta pesquisa é classificada como qualitativa de natureza aplicada, e foi dividida em duas etapas. Na primeira foi elaborada uma pesquisa bibliográfica de caráter exploratório, que proporcionou o embasamento teórico para fundamentar o estudo e levantar os principais problemas relacionados à tarefa de recuperar informações. Na sequência, a pesquisa aplicada consistiu na proposição do método em resposta à problemática identificada. Dentre os principais resultados está a proposição do método de contextualização e expansão de consultas e o desenvolvimento de um software denominado ContextOnSearch, um mecanismo de busca Web com uma interface baseada em uma caixa de texto livre que implementa o método proposto. Para a realização dos testes foi criada uma coleção composta por 481 documentos oriundos do Jornal de Pediatria e publicados entre os anos 2016 e 2020. Foi utilizada ainda uma ontologia da área biomédica denominada *Pediatric Terminology*. Os resultados indicam um aumento da revocação sem perdas significativas na precisão e uma melhoria na classificação pela relevância dos resultados. Conclui-se que o uso de ferramentas de apoio ao usuário em mecanismos de busca pode facilitar a formulação de expressões de busca e possibilitar melhorias na comunicação entre usuários e sistemas, alcançado resultados mais relevantes e contribuindo com o processo de recuperação de informação.

Palavras-chave: Recuperação da Informação. Ontologias de domínio. Contextualização da expressão de busca. Expansão de consultas. ContextOnSearch.

ABSTRACT

An information retrieval system is made up of three basic elements: the documents representations, a user's search expression and some form of comparison between these two elements. On one hand, the collection is composed a moment before the searches, and each document can be represented using automated techniques. On the other hand, the user's need for information is only noticed after its enunciation through a query. The elaboration of a search expression that accurately represents a user's information needs can be a complex task. In this sense, ontologies in the role of vocabulary control instruments can be used to improve this task. Ontologies allow, the contextualization of the search, among other functions, using its terminological structure to search for the terms/concepts that make up the query, it is possible to determine the context of the search. From this context, which was selected by the user and represented by an ontology, the query can be expanded using related terms/concepts. Thus, this research aims to propose an interactive method of contextualizing the information needs and query expansion in information retrieval systems using the terminological structure of domain ontologies. As specific objectives, the thesis proposes to: a) discuss the Information Retrieval process, highlighting the relationship between search expression and retrieved results; b) discuss the concepts and characteristics of ontologies, exploring their usefulness in the processes of information retrieval, expansion and query contextualization; c) develop a method of search contextualization, through the terms that make up the query and from that the query expansion from the identification of related concepts (generic, specific and equivalent) to an initial concept, using domain ontologies for both. d) implement a prototype of a Web information retrieval system to demonstrate the use of the proposed method in a controlled environment; e) analyze the results obtained in relation to relevance of the query. This research is classified as qualitative of an applied nature, and it was divided into two stages. In the first one, an exploratory bibliographical research was carried out, which provided the theoretical basis to support the study and raise the main problems related to the task of retrieving information. Afterwards, the applied research consisted of the proposal of the method in response to the identified problem. Among the main results is the proposition of the method of contextualization and query expansion and the development of a software called ContextOnSearch, a Web search engine with an interface based on a free text box that implements the proposed method. A collection of 481 documents from the Jornal de Pediatria and published between 2016 and 2020 was created to carry out the tests. An ontology from the biomedical area called Pediatric Terminology was also used. The results indicate an increase in recall without significant losses in precision and an improvement in ranking by the relevance of the results. It is concluded that the use of user support tools in search engines can facilitate the formulation of search expressions and enable improvements in communication between users and systems, achieving more relevant results and contributing to the information retrieval process.

Keywords: Information Retrieval; Domain Ontologies; Search Expression Contextualization; Query Expansion; ContextOnSearch.

LISTA DE FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

ACM	*Association for Computing Machinery*
API	*Application Programming Interfaces*
BDTD	Biblioteca Digital Brasileira de Teses e Dissertações
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CI	Ciência da Informação
CSS	*Cascading Style Sheets*
CSV	*Comma-separated values*
DOAJ	*Directory of Open Access Journals*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
IBGE	Instituto Brasileiro de Geografia e Estatística
IDF	*Inverse Document Frequency*
IEEE	*Institute of Electrical and Electronic Engineers*
JS	*JavaScript*
JSON	*JavaScript Object Notation*
MEDLARS	*Medical Literature Analysis and Retrieval System*
NICHD	*National Institute of Child Health and Human Development*
OWL	*Web Ontology Language*
OWL DL	*Web Ontology Language Description Logic*
PEDTERM	*Pediatric Terminology*
PHP	*PHP: Hypertext Preprocessor*
REST	*Representational State Transfer*

RDF	*Resource Description Framework*
RDFS	*RDF Schema*
RI	Recuperação de Informação
RII	Recuperação Inteligente da Informação
SCIELO	*Scientific Electronic Library Online*
SOC	Sistemas de Organização do Conhecimento
SPARQL	*Protocol and RDF Query Language*
SQL	*Structured Query Language*
SRI	Sistemas de Recuperação de Informação
TF	*Term Frequency*
TREC	*Text REtrieval Conference*
URL	*Uniform Resource Locator*
URI	*Uniform Resource Identifier*
WWW	*World Wide Web*
W3C	*World Wide Web Consortium*
XML	*Extensible Markup Language*

SUMÁRIO

1 Introdução

A Ciência da Informação (CI) é uma das áreas que se concentra em estudar o fluxo das informações, desde sua criação até o momento de sua utilização. No passado, os estudos sobre a informação repousavam sobre as áreas da documentação, biblioteconomia e arquivologia (NAKANO, 2019). No entanto, a partir da revolução da informação, iniciada em meados de 1947 com o advento do computador (CAVALCANTI, 1995) a área passou a ter um caráter muito mais interdisciplinar.

De acordo com Schiessl (2015, p. 27),

> A capacidade da sociedade contemporânea de armazenar, organizar, selecionar e compreender a informação disponível está inexoravelmente vinculada às tecnologias. Das estantes de bibliotecas aos incontáveis repositórios digitais espalhados pelo mundo, nunca foi tão difícil encontrar a informação adequada. Paradoxalmente, jamais houve tanta facilidade em acessá-la. Virtualmente não há barreiras, quase tudo está a um clique no mouse de computadores pessoais que vasculham fontes disponíveis na *World Wide Web*, ou simplesmente Web.

Dentre as áreas que estudam e propõem técnicas e teorias para armazenar, organizar, encontrar e acessar informação está a Recuperação de Informação (RI). Um Sistema de Recuperação de Informação (SRI) materializa essas técnicas e teorias, agilizando a busca por de informações a partir de uma coleção de documentos. No ambiente Web, um SRI se apresenta na forma de mecanismo de busca, como os conhecidos Google e Bing.

A recuperação de informação também é o nome de um processo, de uma ação humana que resumidamente consiste em encontrar, a partir de um conjunto de documentos, a informação necessária para responder a uma questão ou resolver um problema. De forma genérica o termo "recuperação de informação" designa a operação pela qual se seleciona documentos de um acervo em função de uma determinada demanda informacional. Implica, portanto, em operar seletivamente um estoque de informação. De acordo com Calvin Mooers, o cientista da computação que cunhou o termo, a Recuperação de Informação é:

> [...] o nome dado ao processo ou método pelo qual um potencial usuário de informação é capaz de converter a sua necessidade de informação em uma lista de citações a documentos em um acervo contendo informações úteis para ele. (MOOERS, 1951, p. 25, tradução nossa).

Um Sistema de Recuperação de Informação (SRI) é formado essencialmente por três elementos: as *representações dos documentos* pertencentes ao acervo; a *expressão de busca* do usuário que representa sua necessidade de informação; e uma *função* ou *mecanismo* capaz de fazer a comparação entre os dois primeiros elementos, resultando uma lista de documentos supostamente úteis para atender a necessidade do usuário (BAEZA-YATES; RIBEIRO-NETO, 2013, p.23).

Os SRIs são intermediários entre a informação registrada e os usuários que a necessitam. Sua eficiência depende da compatibilidade entre a linguagem de representação dos itens de informação e da linguagem de representação da necessidade de informação do usuário. Em um acervo documental, uma informação é recuperada se sua representação coincidir total ou parcialmente com a representação da necessidade do usuário (FERNEDA, 2013). Na maioria dos sistemas, as representações dos documentos e da necessidade do usuário são formalizadas textualmente, sendo, portanto, de natureza linguística.

A representação de um determinado documento inclui os elementos descritivos que o identificam e o caracterizam em um acervo, assim como os elementos indicativos de seu conteúdo informativo, os assuntos por ele tratados. A Ciência da Informação, tendo a informação como seu objeto de pesquisa, possui um conjunto de métodos e metodologias voltadas à representação documental que permitem tornar a informação acessível. Porém, tais metodologias levam em consideração um produto intelectual acabado, geralmente enunciado textualmente, cuja descrição se evidencia por suas características físicas (título, autor, editora, etc) e seu conteúdo informacional pode ser sintetizado/representado por métodos intelectuais ou automatizados.

O processo de recuperação de informação pressupõe a existência de um conjunto ou um acervo de documentos. O acervo documental é constituído *a priori*, anterior a qualquer busca, sendo passível de ser processado por técnicas automatizadas tais como indexação automática, mineração de textos, entre outras. Por outro lado, a necessidade de informação do usuário só é percebida após a sua enunciação por meio de sua expressão de busca. Somente a partir de sua enunciação, a expressão de busca pode ser utilizada em processos interativos que visem a resolução de possíveis ambiguidades e/ou o seu enriquecimento semântico.

A interpretação da necessidade de informação do usuário de um sistema de recuperação de informação é dificultada pelo tamanho (número de termos) reduzido das expressões de busca, não permitindo inferir automaticamente o contexto ou a intenção do usuário. Sendo

assim, diversas técnicas e métodos interativos podem ser utilizados com a finalidade de auxiliar o usuário na tradução de sua necessidade informacional em uma expressão de busca que melhor a represente. Tais métodos assumem importância significativa na medida em que buscam representar com maior fidelidade a real necessidade do usuário.

Nesse sentido, um dos métodos para auxiliar o usuário na enunciação de sua necessidade é a disponibilização de vocabulário adequado, permitindo o acesso à terminologia específica da área ou do domínio de seu interesse. Tais métodos podem ser viabilizados com a utilização de instrumentos de organização do conhecimento, tais como as ontologias.

As ontologias de domínio são essencialmente vocabulários de representação de algum domínio do conhecimento. Elas contêm um conjunto de termos que descreve uma determinada área do conhecimento, incluindo os relacionamentos entre os conceitos e suas definições. São especialmente úteis em processos que envolvam linguagem natural e na comunicação entre humanos e máquinas, favorecendo ainda a interoperabilidade semântica (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999; FARINELLI; ALMEIDA, 2014).

As ontologias possuem ainda estreita relação com a organização do conhecimento, sendo compostas por um conjunto de conceitos e suas relações, sintetizando a forma como as pessoas entendem ou interpretam uma determinada área do saber, e permitindo a representação desse entendimento de maneira formal, ou seja, compreensível por humanos e computadores (MIZOGUCHI, 2004).

É importante ponderar que a afirmação “representação do conhecimento de uma forma compreensível por humanos e computadores” se refere implicitamente a níveis distintos de compreensão, visto que algoritmos de computadores e pessoas compreendem a informação de formas distintas, sendo assim mais adequado afirmar que o conhecimento formalizado pelas ontologias é compreensível por computadores e minimamente compreensível por humanos.

Fachin (2009, p.261) utiliza o termo “Recuperação Inteligente da Informação" (RII) para se referir às abordagens que utilizam ontologias para aprimorar o processo de recuperação de informação. O autor enfatiza a importância de sua utilização nos processos de estruturação e representação da informação, possibilitando com isso uma recuperação mais precisa, explorando o componente semântico das buscas.

Tramontin Junior (2011) insere o contexto como um importante elemento do processo de recuperação de informação. Elemento essencial à compreensão do significado dos termos, o contexto pode ser obtido por meio das palavras que se encontram próximas aos termos em questão. Assim, o contexto ou o sentido de um determinado termo pode ser definido pelo conjunto de outros termos que aparece associado a ele em um domínio do conhecimento. As ontologias são instrumentos que viabilizam esta abordagem, pois contêm a terminologia de uma determinada área do conhecimento, representando seu corpo conceitual. Se um conjunto de termos está contido na ontologia de um determinado domínio, pode-se assumir que tais termos estão de certa forma relacionados a este domínio.

Explorando essa premissa, esta tese tem como contribuição principal a proposição de um método interativo de contextualização e expansão de consultas utilizando ontologias de domínio. O método pode ser utilizado em sistemas de recuperação de informação e mecanismos de busca textuais de propósito geral e tem por objetivo melhorar os resultados da recuperação a partir do aprimoramento da representação das necessidades de informação dos usuários. Seu funcionamento baseia-se em duas etapas: a contextualização da necessidade de informação e a expansão de termos de busca.

A contextualização equivale à identificação do domínio de conhecimento ao qual pertence a necessidade do usuário, e da terminologia específica relacionada à esta área. Seu funcionamento consiste em procurar pelos termos que compõem a expressão de busca no interior das ontologias disponíveis no sistema. As ontologias que possuem termos/conceitos coincidentes com os da busca são candidatas a representar o domínio de conhecimento da consulta. Por meio de uma interface adequada, caberá ao usuário a escolha do contexto que melhor representa a sua busca.

A partir de uma determinada ontologia que representa o contexto selecionado pelo usuário, tem início o processo de expansão de consulta, que consiste em procurar por termos que coincidem com os da consulta inicial. Na sequência são identificados na ontologia os conceitos relacionados oriundos das relações hierárquicas e de equivalência. Por fim apresenta-se em um diagrama hierárquico os conceitos encontrados, possibilitando que o usuário os utilize na reformulação de sua expressão de busca.

1.1 Caracterização do problema de pesquisa

A questão que motivou este estudo foi compreender os fatores externos que interferem no processo de representação da necessidade de informação do usuário durante a enunciação de sua expressão de busca.

Para isso recorreu-se à área de Recuperação de Informação para o entendimento dos processos e dos problemas envolvidos na tarefa de localizar informações. De acordo com Beppler (2008) e Baeza-Yates e Ribeiro Neto (2013), os mecanismos de busca existentes, baseados em uma caixa de texto, apesar de terem se tornado muito populares em função de sua simplicidade de uso, possuem problemas que interferem na tarefa de encontrar as informações desejadas, principalmente em relação à elaboração das expressões de busca.

Pesquisas conduzidas por Camacho (2004), Beppler (2008), Baronas (2011) e Nayak, Prasad e Senapati (2015) apontam que estes problemas podem ser resumidos em:

- dificuldade de expressar necessidades informacionais utilizando palavras-chave (ou termos de busca);
- necessidade de conhecimento sobre o domínio pesquisado, pois quanto maior o conhecimento do usuário sobre o assunto, maior será sua familiaridade com a sua terminologia, e consequentemente ele terá mais condições de enriquecer suas buscas e obter resultados mais relevantes, ao mesmo tempo que, quanto menor seu conhecimento, mais genéricas e inadequadas serão as palavras-chave utilizadas;
- dificuldade em compreender a terminologia e interpretar as relações entre os conceitos de um determinado domínio;
- dificuldade em lidar com os problemas linguísticos como sinonímia, homonímia, paronímia, polissemia, denotação, conotação e outras variações linguísticas que dificultam ainda mais a tradução das necessidades de informação em expressões de busca adequadas;
- falta de mecanismos que auxiliem o usuário identificar termos que podem ser úteis para o aprimoramento da consulta.

Considerando estes problemas, foi feito um levantamento bibliográfico buscando identificar propostas de melhorias e soluções. Com base nos estudos de Jackson e Moulinier

(2007), Woods (2004), Hyvönen *et al*. (2005), Lei, Uren e Motta (2006), Beppler (2008), Baeza-Yates e Ribeiro-Neto (2013) e Shneiderman *et al*. (2017) apresentam-se as seguintes propostas:

- disponibilizar componentes visuais que auxiliem na elaboração das expressões de busca;
- disponibilizar o vocabulário adequado para permitir a elaboração de expressões de busca mais significativas e melhor contextualizadas, elaboradas a partir dos conceitos definidos na ontologia;
- apresentar por meio de formatos gráficos os conceitos e instâncias de um domínio para possibilitar melhor entendimento;
- possibilitar a formulação de consultas menos ambíguas, visto que os usuários terão acesso a terminologia específica do domínio;
- possibilitar a descoberta de novos termos por meio da apresentação de termos mais gerais e específicos obtidos a partir das relações presentes na ontologia.

É possível notar em todas as sugestões um aspecto em comum: a necessidade de recursos e/ou ferramentas que facilitem a elaboração das expressões de busca, favorecendo a enunciação das necessidades informacionais do usuário e por consequência a recuperação de informação relevante.

Nesse sentido, as estruturas de conhecimento capazes de representar os conceitos e as relações de um determinado domínio tornam-se imprescindíveis. As ontologias são ótimas candidatas para utilização em tais abordagens, pois "incluem regras de inferências e axiomas que não existem em nenhum outro tipo de sistema de organização do conhecimento" (CARLAN, 2010, p. 36).

Fundamentada nesta problemática, a proposta desta pesquisa busca descrever um método de contextualização e expansão da expressão de busca baseado em ontologias, promovendo a materialização de algumas das melhorias supracitadas nos mecanismos de busca.

Durante o levantamento bibliográfico foi possível notar ainda, a partir dos estudos de Lai e Soh (2004) e Allan, Carterette e Lewis (2005), que quando se busca informação utilizando SRI os usuários tem dificuldades que resultam no gasto de um tempo considerável na tentativa de localizar a informação requerida. Ao passo que Beppler (2008) afirma que um dos problemas mais comuns está na elaboração inadequada das consultas e na falta de interatividade dos sistemas de busca.

Marchionini (1989), Kuhlthau (2003), Sonnenwald, Wildemuth e Harmon (2001), Capra e Pérez-Quiñones (2005) corroboram a percepção de que os usuários enfrentam problemas ao descrever suas necessidades, explicando que essa inabilidade em expressar precisamente suas necessidades informacionais faz com que eles entrem em uma sequência indeterminada de interações com o sistema de busca.

Desta forma, a tarefa de elaborar uma expressão de busca composta por um ou mais termos que representem as necessidades de informação muitas vezes esbarra na falta de conhecimento sobre o assunto. Beppler (2008) explica que esta dificuldade está diretamente relacionada com a experiência do usuário e com seu entendimento do problema. O autor ainda salienta que durante o processo de busca, a cada interação com o sistema, o usuário acessa novas informações que afetam o entendimento de suas necessidades e, consequentemente, aquilo que ele está buscando.

Há ainda outros aspectos a se considerar, como o fato de que o usuário nem sempre percebe qual é a informação de que necessita, ou seja, não reconhece suas necessidades informacionais (FORSETLUND; BJORNDAL, 2001; TAYLOR, 2015; WILLIAMSON *et al.*, 1989). Portanto, o processo de busca por informação possui natureza dinâmica, e mesmo sendo uma tarefa frequente no cotidiano das pessoas ainda demanda uma quantidade significativa de tempo e esforço.

Com base nas propostas de melhorias nos sistemas de recuperação de informação entende-se que combinar a simplicidade dos mecanismos de busca baseados em caixas de texto, com abordagens que auxiliem os usuários na elaboração de expressões de busca mais adequadas frente à suas necessidades informacionais é um caminho oportuno.

Nesse sentido, um método que faça uso das relações semânticas formalizadas nas ontologias para expansão dos termos da expressão de busca pode contribuir significativamente com o processo de recuperação de informação. Pois, a contextualização da expressão de busca e a sugestão de termos relacionados pode resultar em uma técnica promissora no sentido de facilitar a localização de informações relevantes.

O problema central desta pesquisa é a dificuldade dos usuários em representar suas necessidades informacionais em uma expressão de busca. Sendo assim, busca-se colaborar com o desenvolvimento de abordagens que possam aprimorar os sistemas de recuperação de informação textuais no contexto da Web, propondo um método de torná-los precisos e interativos.

O diferencial deste estudo é a utilização de ontologias de domínio para contextualização e expansão das consultas, possibilitando com isso o fornecimento de informações adicionais que podem ser visualizadas a cada instância recuperada, permitindo ao usuário utilizá-las para refinar ou formular novas consultas.

1.2 Hipóteses

A partir da caracterização do problema da pesquisa, foram identificadas as seguintes hipóteses norteadoras:

- a estrutura terminológica de uma ontologia de domínio pode ser utilizada como elemento capaz de contextualizar a consulta do usuário de um sistema de recuperação de informação;
- as relações semânticas entre os conceitos de uma ontologia permitem efetuar a expansão da expressão de busca inicialmente formulada pelo usuário, agregando a ela novos termos a fim de melhorar a eficiência da recuperação;
- a apresentação da hierarquia de conceitos relacionados à expressão de busca do usuário favorece a compreensão da terminologia de um domínio, bem como explicita as relações entre conceitos, permitindo que o usuário selecione os termos que serão utilizados na expansão da busca inicialmente formulada.

1.3 Proposição

Considerando que os mecanismos de busca Web mais populares utilizam caixas de texto livre para que os usuários interajam com o sistema, parte-se do pressuposto que eles não oferecem formas de auxílio significativas à etapa de elaboração da expressão de busca, deixando sob a responsabilidade dos utilizadores a representação adequada das necessidades informacionais.

Desta forma, esta pesquisa se propõe a utilizar ontologias de domínio e os relacionamentos semânticos entre seus conceitos para contextualizar as expressões de busca e identificar termos relacionados para sua expansão, facilitando e aprimorando assim a etapa de elaboração da consulta.

1.4 Tese

A tese desta pesquisa enuncia-se como:

A estrutura terminológica definida por uma ontologia de domínio pode ser utilizada na contextualização e expansão de consultas, permitindo uma melhor representação das necessidades informacionais dos usuários e refletindo diretamente na eficiência de um sistema de recuperação de informação.

1.5 Objetivo geral

Esta pesquisa tem como objetivo geral propor um método de contextualização e expansão de consultas em sistemas de recuperação de informação utilizando a estrutura terminológica definida por uma ontologia de domínio.

1.6 Objetivos específicos

a) Discutir o processo de Recuperação de Informação, evidenciando a relação entre a expressão e busca e os resultados recuperados;

b) Discutir os conceitos e características das ontologias, explorando sua utilidade nos processos de recuperação de informação, expansão e contextualização das consultas;

c) Desenvolver um método de utilização de ontologias de domínio para:

— contextualização da consulta, por meio dos termos que a compõem;

— identificação de conceitos relacionados (genéricos, específicos e equivalentes) à um conceito inicial e, utilizá-los para expansão de consultas;

d) Implementar um protótipo de um sistema de recuperação de informação Web para demonstrar a utilização do método proposto em um ambiente controlado;

e) Analisar os resultados obtidos em relação à relevância com a expressão de busca.

1.7 Justificativa

Pesquisas na área de Recuperação de Informação são imprescindíveis à evolução tecnológica da sociedade, em específico, a temática da representação das necessidades informacionais dos usuários em SRI pode promover benefícios significativos, pois discute e reflete sobre a dinâmica de interação dos usuários com os buscadores e sobre a efetividade na localização de informações.

No contexto social, a proposição de abordagens que facilitem a localização de informações relevantes, demandando menos tempo e esforço, tem o potencial disseminador de conhecimento para uma sociedade pois, ao tornar os sistemas de busca mais interativos e eficientes na localização das informações, possibilita-se que mais pessoas encontrem respostas aos seus questionamentos, que por sua vez, possam produzir soluções para problemas e com isso gerar novos conhecimentos.

No âmbito técnico e profissional, a pesquisa pode fornecer subsídios para o desenvolvimento de sistemas de recuperação de informação mais adequados para o atendimento das necessidades dos usuários. O método evidencia a necessidade de considerar as dificuldades do usuário como barreias à localização de informações, propondo um mecanismo de auxiliar o usuário na tarefa de construção da expressão de busca.

No âmbito científico este estudo contribuí com a área de RI, por meio da proposição do método de contextualização e expansão de consultas, que auxilia os usuários na elaboração da expressão de busca, representando mais adequadamente suas necessidades informacionais. Contribui-se ainda com a evolução dos SRI, que com expressões de busca mais claras e precisas são capazes de recuperar informações mais relevantes.

Baseado no que foi discutido na caracterização do problema de pesquisa, este estudo considera que os mecanismos de busca necessitam de recursos de auxílio aos usuários na elaboração de expressões de busca, promovendo a superação de dificuldades como a falta de conhecimento da terminologia e do domínio de interesse.

Desta forma o método proposto implementa algumas soluções para estes problemas, como o uso de componentes visuais que apresentam a terminologia do domínio e as relações entre os conceitos, organizadas em formato hierárquico. Evidencia-se desta forma o caráter inovador e benéfico deste estudo.

1.8 Metodologia

As pesquisas científicas são realizadas com rigor, ética, procedimentos metodológicos e matrizes teóricas específicas. Desenvolvidas por profissionais de diferentes áreas do conhecimento elas tem como propósito responder, de forma profunda e sistemática, aos questionamentos que emergem do contexto profissional (DEL-MASSO; COTTA; SANTOS, 2014).

Fleury e Werlang (2017, p.11) afirmam que

> [...] a pesquisa científica não trata apenas de resenhas bibliográficas ou elucubrações genéricas. Ela visa produzir conhecimento por meio de conceitos, tipologias, verificação de hipóteses e elaboração de teorias que possuam relevância na disciplina acadêmica ancoradas de determinadas escolas de pensamento.

De acordo com Severino (2007), as pesquisas científicas podem ser classificadas quanto à natureza (básica ou aplicada), ao tipo (bibliográfica, documental, experimental, exploratória, descritiva, entre outras), e quanto à abordagem (qualitativa ou quantitativa).

Quanto à natureza, essa pesquisa é classificada como aplicada, que segundo Appolinário (2011, p. 146), é realizada com o intuito de "[...] resolver problemas ou necessidades concretas e imediatas".

Em uma definição mais detalhada Thiollent (2018) explica que a pesquisa aplicada se concentra em torno dos problemas presentes nas atividades das instituições, organizações, grupos ou atores sociais. Ela está empenhada na elaboração de diagnósticos, identificação de problemas e busca de soluções que respondem a uma demanda específica.

Em relação ao tipo, entende-se que a pesquisa é bibliográfica de caráter exploratório, pois são necessários embasamentos teóricos para a definição do problema de pesquisa e formulação das hipóteses.

Fundamentalmente a pesquisa bibliográfica é aquela que se realiza a partir do

> [...] registro disponível, decorrente de pesquisas anteriores, em documentos impressos, como livros, artigos, teses etc. Utilizam-se dados de categorias teóricas já trabalhadas por outros pesquisadores e devidamente registrados. Os textos tornam-se fontes dos temas a serem pesquisados. O pesquisador trabalha a partir de contribuições dos autores dos estudos analíticos constantes dos textos (SEVERINO, 2007, p.122).

Para Gil (2008, p.37-38), os estudos bibliográficos exploratórios devem proporcionar uma visão geral sobre os temas, explanando conceitos essenciais apresentados por autores relevantes nas respectivas áreas e familiarizando o leitor com os conhecimentos essenciais à compreensão de uma proposta.

Na pesquisa bibliográfica, que ocorreu durante todo o desenvolvimento desta tese, foram analisadas as referências teóricas extraídas de artigos, teses, dissertações e livros que versam sobre a Recuperação da Informação e Ontologias, que representam as bases teóricas

deste estudo. Foram consultadas as seguintes bases de dados eletrônicas: ACM, Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), *Directory of Open Access Journals* (DOAJ), Google Acadêmico, IEEE, Periódicos Capes, *Scientific Electronic Library Online* (Scielo), *Scopus* e *Web of Science*. Tais fontes foram escolhidas por indexarem diversos periódicos nacionais e internacionais de caráter multidisciplinar.

As palavras chave utilizadas para as buscas foram "recuperação de informação" "ontologias" e suas respectivas traduções para o idioma inglês "*information retrieval*" e "*ontology*". Em um segundo momento foi feita uma filtragem dos trabalhos, selecionando apenas os que tratavam dos sistemas de recuperação de informação e do usuário. A terceira etapa foi uma análise de conteúdo, mantendo apenas aqueles cujo objeto de interesse do estudo fosse a representação das necessidades do usuário (por meio da expressão de busca). Foram descartados ainda trabalhos com mais de 20 anos, limitando a pesquisa entre os anos 2000 e 2020. Este levantamento bibliográfico resultou em um total de 33 trabalhos.

Partindo do problema delimitado na fundamentação teórica, propôs-se um método de contextualização e expansão de consultas utilizando ontologias de domínio, cujo objetivo é auxiliar os usuários a representarem suas necessidades informacionais durante o processo de elaboração das expressões de busca. A proposta relaciona-se ainda com as áreas da teoria da comunicação, teoria do conceito, linguística, terminologia e expansão de consultas, que devido à sua interdisciplinaridade com a área de RI abordam conceitos necessários à compreensão do método proposto.

Severino (2007) considera mais adequado empregar os termos abordagem qualitativa e abordagem quantitativa, por considerar que muitas são as pesquisas com metodologias diferenciadas, as quais podem caracterizar-se como uma abordagem qualitativa e abordagem quantitativa. É importante lembrar que "são várias as metodologias de pesquisa que podem adotar uma abordagem quantitativa. Modo de dizer que faz referência mais a seus fundamentos epistemológicos do que propriamente a especificidades metodológicas" (SEVERINO, 2007, p. 119).

Sendo assim, a abordagem deste estudo é qualitativa, pois visa analisar os resultados obtidos da utilização do método de contextualização em expansão de consultas em um sistema de busca Web. Para o desenvolvimento desta etapa foi implementado um protótipo de *software* denominado ContextOnSearch, cujo propósito metodológico foi demonstrar o funcionamento do método proposto. É importante salientar que não foi feita uma validação com usuários, sendo

que as análises e ponderações apresentadas ao longo do trabalho decorrem de percepções pessoais obtidas a partir da utilização do *software* para a execução de testes práticos.

1.9 Terminologia adotada

O tema central desta pesquisa, a Recuperação da Informação, permeia os campos científicos da Ciência da Informação e da Ciência da Computação. Surgem daí problemas terminológicos decorrentes de diferentes termos utilizados para referenciar um mesmo conceito, além de traduções oriundas do idioma inglês.

Neste trabalho considera-se uma equivalência entre os termos “documento” e “informação”. Parte-se da definição de Buckland (1991) que considera informação como coisa, englobando objetos, documentos, textos ou eventos, cuja materialidade pode ser observada.

Quanto aos documentos, desconsiderando a complexidade inerente à sua definição, adota-se a concepção de Pinto Molina, García Marco e Agustin Lacruz (2002) *apud* Torres e Almeida (2014, p. 3)

> [...] o documento cumpre várias funções diferentes, atuando como ferramenta de comunicação e meio de expressão, proporciona informação sobre a realidade, é um instrumento de controle, uma ferramenta cognitiva que ajuda a pensar de forma mais eficaz. Serve ainda, como memória externa, permitindo a construção de uma memória socialmente compartilhada, que constitui um instrumento para a construção da cultura.

Desta forma, os documentos podem ser compreendidos como suportes da informação registrada, e, no contexto dos mecanismos de busca recupera-se informação ao recuperar os documentos que as contém.

Outra reflexão se dá pela utilização de "sistema de recuperação de informação" como um termo genérico para referenciar sistemas implementados em qualquer tipo de plataforma. Por conseguinte, o termo "sistema de recuperação de informação Web" é utilizado para tratar dos sistemas desenvolvidos especificamente para esse ambiente. Os termos “motor de busca”, “mecanismo de busca”, “ferramenta de busca” e “buscador” serão considerados sinônimos.

Os termos “expressão de busca” e “consulta” são tradicionalmente encontrados nas literaturas da Ciência da Informação e da Computação, respectivamente. Por extensão, o conjunto formado por uma ou mais palavras que representa linguisticamente uma necessidade de informação é denominado "termos de busca" ou "termos da consulta".

A Teoria do Conceito de Dahlberg (1978) entende um conceito como uma unidade de conhecimento, enquanto Barros (2016) explica que a comunicação em um determinado campo de saber necessita da comunicação de conceitos, que por sua vez demandam algum tipo de suporte para sua materialização, que, na linguagem verbal podem ser representados pelos termos, que remetem à ideia relacionada a um conceito. Sendo assim, utiliza-se no decorrer do texto o vocábulo "termo" no sentido (ou como sinônimo) de "conceito".

A última ponderação refere-se ao uso da palavra "domínio" para se referir ao "contexto" de um determinado termo. A definição de contexto pode variar em razão de fatores como a área de pesquisa e objeto de análise. No entanto, este trabalho limita-se a considerar o "contexto" como o domínio de utilização de um termo, sua modalidade, condições de utilização e elementos semânticos e definitórios que permitem estabelecer a relação entre um termo e o conceito por ele representado (CERVANTES, 2020).

A autora ainda explica que o contexto exprime a ideia completa sobre um determinado termo, remetendo ao seu significado ou tema em uma determinada situação. Sendo assim adota-se do decorrer do estudo uma equivalência semântica entre os vocábulos "domínio" e "contexto".

1.10 Organização do trabalho

Esta tese está organizada em seis capítulos, começando pela **Introdução**, que apresenta uma narrativa que explicita a linha de pensamento do autor em relação aos conceitos abordados. Um delineamento da pesquisa é apresentado na sequência seguido pela identificação do tema, uma caracterização do problema, as hipóteses levantadas, a proposição e a tese, os objetivos gerais e específicos e os procedimentos metodológicos adotados para a execução do trabalho.

No segundo capítulo, denominado **Recuperação de Informação**, é apresentada a área homônima, onde aborda-se brevemente os aspectos históricos, as definições, o processo de recuperação de informação e os modelos clássicos, objetivando compreender os aspectos que influenciam no desempenho dos sistemas de recuperação de informação. Por fim, aborda-se a RI no contexto da Web, que possui características únicas.

O terceiro capítulo discorre sobre as **Ontologias**, apresentando uma revisão de literatura que passa pelas definições do termo, componentes e características. Por fim são discutidas algumas das motivações para sua utilização no contexto da recuperação da informação e a linguagem para representação do conhecimento OWL.

O quarto capítulo apresenta o **Método de contextualização e expansão de consultas baseado em ontologias de domínio**, o cerne desta pesquisa. São abordados os detalhes de como as ontologias de domínio podem ser utilizadas para os processos de contextualização e expansão das consultas. Apresenta-se brevemente ainda alguns conceitos da área da comunicação e a teoria do conhecimento, concernentes à compreensão do método.

No quinto capitulo é apresentado o **ContextOnSearch: um sistema de recuperação de informação com recursos de contextualização e expansão de consultas**, que é um *software* desenvolvido para demonstrar a implementação do método proposto em um mecanismo de busca. São abordados os aspectos de seu desenvolvimento, as tecnologias empregadas e alguns resultados obtidos.

O sexto e último capítulo contém as **Considerações Finais**, onde são apresentadas algumas discussões à cerca dos resultados obtidos seguidas de reflexões, limitações e contribuições do trabalho terminando com sugestões para pesquisas futuras.

2 Recuperação de Informação

Este capítulo apresenta uma revisão de literatura sobre a área de Recuperação de Informação, um dos assuntos basilares desta pesquisa. Inicialmente são abordados os aspectos históricos do termo e suas definições mais recorrentes. Na sequência discute-se o processo, seus componentes e modelos. Por fim aborda-se a recuperação de informação no ambiente Web, um contexto que possui particularidades e desafios únicos que possui uma dinâmica distinta das bibliotecas, repositórios ou centros de documentação.

O termo Recuperação de Informação (RI) é a tradução de *Information Retrieval* (*IR*), que foi apresentado em 1951 pelo cientista da computação Calvin Mooers. Ele a define como "a disciplina que trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação" (MOOERS, 1951, p.25, tradução nossa).

Baeza-Yates e Ribeiro-Neto (2013) destacam que alguns pesquisadores tiveram contribuições significativas para a área, tais como Allan Kent e seus colegas, que em 1955 publicaram um artigo descrevendo as métricas de precisão e revocação, bastante conhecidas e utilizadas para a avaliação de sistemas de recuperação de informação (KENT *et al.*, 1955). O projeto Cranfield idealizado por Cyril Cleverdon e sua equipe e a avaliação do Sistema *Medical Literature Analysis and Retrieval System* (MEDLARS) levada a efeito por Frederick Wilfrid Lancaster colaboraram com a com a compreensão dos fatores que afetam o desempenho dos sistemas de recuperação de informação (CLEVERDON, 1967; LANCASTER, 1968). Por fim, ressaltam-se as pesquisas de Gerard Salton (SALTON, 1968) e Karen Sparck-Jones (SPARK-JONES, 1972), que culminaram no desenvolvimento de conceitos fundamentais, como o ranqueamento de documentos e os mecanismos de busca.

Mooers (1951) afirma que a recuperação de informação é um processo crucial para documentação e organização do conhecimento. No entanto, é importante destacar que os conceitos de "informação" e "conhecimento" são passíveis de conceituações distintas, e não há um consenso entre pesquisadores quando às suas definições. Neste trabalho adota-se o entendimento de Buckland (1991) e Le Coadic (2004) em que a informação é entendida como

algo palpável, tangível, inscrita em algum suporte e que transmite uma mensagem ao indivíduo que é capaz de compreender a linguagem ao qual a informação está inscrita.

A Recuperação da Informação é uma área de pesquisa interdisciplinar, que recebe aportes de outras áreas como a Ciência da Computação, por meio de estudos que melhoram o desempenho dos sistemas de recuperação de informação, reduzindo seu tempo de resposta, aumentando sua escalabilidade e da Interação Humano-Computador (IHC), que se dedica a estudar a interação entre as pessoas e os computadores.

A Ciência da Informação possui um olhar mais voltado às informações, buscando compreender as dinâmicas e os fluxos informacionais e propondo formas de organização e representação das informações, e com isso colaborando também melhorias com no processo de RI como um todo.

Para compreender o termo “Recuperação de Informação” como área de pesquisa e também como processo buscou-se conceituações de pesquisadores ao longo dos anos, começando por Salton (1968) que a define como “uma área de pesquisa que tem por premissa básica a estrutura, análise, organização, armazenamento, recuperação e busca da informação relevante”. Rijsbergen (1979, p.3, tradução nossa) a define como “o processo de recuperação de documentos que sejam provavelmente relevantes para a necessidade de informação do usuário, expressa por meio de uma busca”.

Ingwersen (1992, p.49, tradução nossa) entende que “a recuperação de informação está relacionada com os processos vinculados com a representação, armazenamento, busca e identificação da informação relevante para a necessidade de informação de um usuário humano”. Tem por objetivo investigar e compreender os processos de recuperação de informação para desenhar, construir e avaliar SRI para que o processo de comunicação entre homens e máquinas seja eficaz e facilitado.

A partir das conceituações apresentadas é possível notar que a Recuperação de Informação é composta por etapas, que resumidamente são: a construção da representação dos documentos de um acervo, uma expressão de busca que representa a necessidade de informação do usuário, um mecanismo que faz a comparação destas duas e, por fim, a apresentação dos resultados supostamente mais relevantes frente a esta necessidade.

É correto afirmar ainda que a RI é um processo, composto por elementos que possuem funções bem definidas e que influenciam em seu resultado final, a recuperação relevante de informação para um usuário.

Analisando as definições de Salton (1968), Rijsbergen (1979) e Ingwersen (1992), é perceptível o aumento gradativo da importância do usuário, pois na primeira definição, não há menção ao usuário, enquanto na segunda, o autor afirma que o objetivo da RI é atender a uma necessidade do usuário. Porém, é só na terceira definição, distante mais de 20 anos da primeira, que fica claro que o objetivo das investigações é atender as necessidades informacionais dos usuários de forma fácil e eficiente.

Considerar o usuário como elemento central no processo de recuperação de informação é imprescindível, pois desconsiderá-lo incorreria em sistemas e mecanismos ineficientes, incapazes de estabelecer uma comunicação e uma aproximação de seus utilizadores, possibilitando o acesso à informação sobre a necessidade de informação, pode-se construir expressões de busca mais satisfatórias.

Spink e Saracevic (1993, grifo nosso) definem a recuperação de informação como:

> Um processo de alta complexidade envolvendo numerosos fatores e variáveis além de decisões e o entrelaçamento dos subprocessos inter-relacionados com a busca. Uma das chaves para um dos subprocessos é a da **seleção de termos para a estratégia de busca** que por sua vez é influenciada por outros fatores, particularmente os relacionados com os resultados. Sintetizam essa questão com uma pergunta para as investigações: que termos de busca devem ser selecionados para um determinado tema que represente efetivamente o problema de informação do usuário?

Nesta definição observa-se a compreensão dos autores que o processo de RI é complexo, pois sofre influência de vários fatores. Além disso, o destaque da escolha dos termos na estratégia de busca demonstra uma clara preocupação com a seleção dos termos mais adequados para representar a necessidade de informação do usuário. Esta preocupação foi uma das fontes de inspiração para a idealização do método proposto neste trabalho.

Avançando para definições mais recentes, Büttcher, Clarke e Cormack (2010, p.1, tradução nossa) que entendem que a Recuperação de Informação

> se preocupa em representar, pesquisar e manipular grandes coleções de textos e registros produzidos pelos humanos. Sistemas de recuperação de informação estão disseminados a ponto de milhares de pessoas dependerem deles para desempenhar suas atividades de trabalho, ensino ou entretenimento. Mecanismos de busca Web como Google, Bing e outros são de longe os mais

> populares serviços de recuperação de informação, provendo acesso a informações atualizadas sobre documentos técnicos, localização de pessoas e organizações, notícias, eventos e comparações para compras em geral.

Na definição destes autores pode-se perceber a menção ao volume de informações, a preocupação de considerar como a grande quantidade de registros disponíveis influencia no processo de localizar informações. Nota-se ainda a menção aos sistemas de recuperação de informação na Web, consubstanciados em mecanismos de busca, serviços de suma importância no cotidiano das pessoas.

Baeza-Yates e Ribeiro-Neto (2013, p.1) explicam que:

> A recuperação de informação é responsável pela representação, armazenamento, organização e acesso a itens de informação, como documentos textuais e multimídia, páginas web, registros estruturados e semiestruturados, com o objetivo de fornecer aos usuários facilidade de acesso as informações de seu interesse.

Os autores destacam a composição heterogênea das coleções, compostas por documentos textuais e multimídia nas páginas Web, registros que representam um volume significativo dos acervos atuais e que possuem uma natureza distinta de outros documentos como livros e textos acadêmicos, tornando-se assim fatores importantes a serem considerados no processo de recuperação de informação.

Em termos de pesquisa, a área de RI pode ser estudada sob dois pontos de vista distintos e complementares: um centrado nos sistemas e outro centrado no usuário. No primeiro, a pesquisa consiste principalmente na construção de índices eficientes, no processamento de consultas com alto desempenho e no desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados. Na visão centrada no usuário, a pesquisa concentra-se em estudar o comportamento do usuário, entender suas principais necessidades e determinar como esse entendimento afeta a organização e a operação do sistema de recuperação (BAEZA-YATES; RIBEIRO-NETO, 2013, p.1).

Essa conceituação sobre as distintas abordagens da área relaciona-se com a proposta dos paradigmas de Capurro (2003): o Físico, o Cognitivo e o Social, cuja diferenciação é dada por Almeida *et al.* (2007, p.25):

> O que diferencia o enfoque de cada um dos paradigmas são as diferentes abordagens de fornecimento de informações, sendo que no Paradigma Físico busca-se utilizar informações para alimentar sistemas computacionais, o Paradigma Cognitivo leva em consideração as informações que satisfaçam as

> necessidades individuais de cada indivíduo mediante o seu processo mental e o Paradigma Social considera as informações de acordo com o contexto social ao qual o usuário pertence.

É importante frisar que ambas as visões, a centrada no usuário e a centrada nos sistemas não são isoladas, elas se complementam e são interdependentes, de forma que aprimoramentos em qualquer uma delas implicam em melhorias no processo de RI como um todo.

Para gerenciar as informações são utilizados os Sistemas de Informação, que Buckland (1991, tradução nossa) define como "quaisquer unidades que coletem, organizem e disponibilizem objetos potencialmente informativos". No entanto, a expressão "Sistemas de Informação" é demasiado genérica, pois pode remeter a um conceito da área da computação, que define um sistema de informação como um conjunto de elementos interconectados que incluem uma entrada de dados, um mecanismo de processamento e conversão dessa entrada em uma saída da informação (GONÇALVES, 2011, p. 50).

A área de RI estuda especificamente os Sistemas de Recuperação de Informação (SRI), que podem ser definidos como "um conjunto de ações que permitem a organização e a comunicação da informação documentária, ou seja, possibilitam a transferência da informação documentária através de procedimentos seletivos que regulam sua geração, distribuição e uso." (LIMA, 2004, p. 96).

Uma definição clássica e detalhada de SRI é dada por Cesarino (1985, p. 157):

> Os sistemas de recuperação da informação podem ser definidos como um conjunto de operações consecutivas executadas para localizar, dentro da totalidade e informações disponíveis, aquelas realmente relevantes. Para isso, executam as funções de seleção, análise, indexação e busca das informações. Em todas essas etapas a interação usuário x sistema é fundamental [...]

Para Ferneda e Dias (2013, p.2) um SRI pode ser entendido como um elemento intermediário entre as informações registradas nos documentos armazenados em algum acervo, e os usuários com suas necessidades informacionais. Sua eficiência depende essencialmente da combinação entre a linguagem utilizada na representação dos itens de informação e nas buscas dos usuários.

2.1 O processo de recuperação de informação

O processo de recuperar informações pode ser entendido como a ação de localizar documentos supostamente úteis à necessidade informacional de um usuário. Sendo assim em seu formato mínimo ele é composto por:

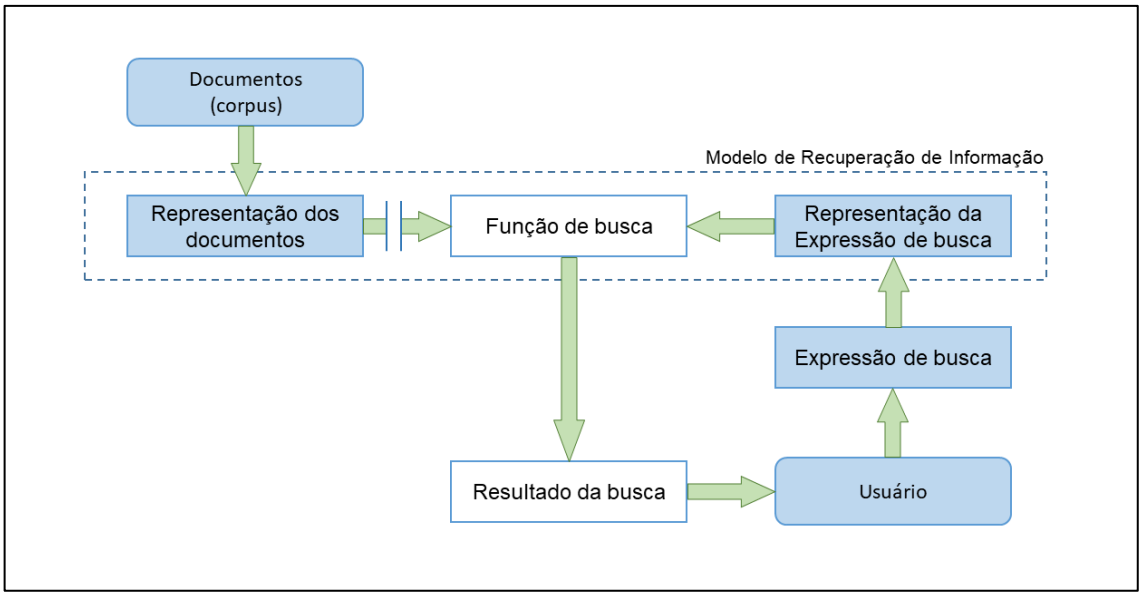
a) uma **expressão de busca** proveniente de uma necessidade de informação de um usuário;

b) uma **representação dos documentos** pertencentes ao acervo. Um exemplo de uma representação é um índice de um livro, que é uma representação sucinta do conteúdo do mesmo, criada para facilitar o acesso ao mesmo;

c) uma **função** ou mecanismo capaz de fazer a comparação das duas anteriores, possibilitando assim a criação de uma lista de registros úteis ao usuário.

É importante observar que o processo de RI pode ser ligeiramente diferente dependendo do pesquisador que o apresenta, por isso optou-se por trazer também a representação sob o olhar de Ferneda (2012, 2013), que pode ser visto na Figura 1. Para o autor o processo possui os mesmos elementos essenciais, com o diferencial de um detalhamento maior do fluxo de execução das ações, o que possibilita uma melhor compreensão da influência de cada etapa no processo como um todo.

Figura 1 - Representação do processo de recuperação de informação



Fonte: Adaptado de Ferneda (2013, p.47)

A representação do processo de Recuperação de Informação de Ferneda (2013) é composta por um lado, pelos documentos e suas representações, por outro pelo usuário com suas necessidades informacionais traduzidas em uma expressão de busca e sua respectiva

representação, e em uma posição central a função de busca, responsável pela comparação entre as representações e apresentação das coincidências (a intersecção) entre elas como o resultado da busca. Um detalhe importante apontado pelo autor é que na seta que vai da representação dos documentos para a função de busca há uma secção, isso foi feito para destacar que a construção da representação de um acervo documental, construída por meio de indexação por exemplo, que é feita em momento anterior a execução de qualquer busca, demonstrando assim uma assincronia entre estas ações.

Inicia-se a abordagem dos elementos pertencentes ao processo de RI apresentado na Figura 1 discutindo sobre os **Documentos (corpus)** que são os objetos de informação, os suportes onde o conhecimento é registrado e disponibilizado com objetivo de disseminação e perpetuação. Faz-se importante destacar que não se pode considerar documento apenas objetos palpáveis, pois com o advento dos computadores e seu funcionamento com sinais digitais, há documentos que existem apenas nestes meios, como os *e-books* e os conteúdos multimídia.

Observa-se que este trabalho segue a concepção de Buckland (1991) que compreende o conceito de "informação como coisa" onde o documento é um objeto que contém informações. Um documento pode ser entendido ainda como algo informativo, incluindo por exemplo, objetos, artefatos, imagens e sons (FERNEDA, 2012, p.15).

Le Coadic (2004, p.5) define documento como:

> O termo genérico que designa os objetos portadores de informação. Um documento é todo artefato que representa ou expressa um objeto, uma ideia ou uma informação por meio de signos gráficos e icônicos (palavras, imagens, diagramas, mapas, figuras, símbolos), sonoros e visuais (gravados em suporte de papel ou eletrônico).

A **Representação dos Documentos** ou representação da informação é a etapa que busca descrever o conteúdo dos documentos para que seja possível recuperá-los em momento posterior. Ela consiste essencialmente em identificar os assuntos mais relevantes por meio da análise de conteúdo, traduzir as palavras-chave encontradas utilizando uma linguagem de indexação e por fim armazená-las no SRI para que sirvam de pontos de acesso para este documento.

A análise documentária refere-se ao processo de representação da informação, que pode ser classificado em representação descritiva e representação temática (MAIMONE; SILVEIRA; TÁLAMO, 2011). A primeira descreve as características objetivas do documento, como autor, título e ano da publicação, em seguida classificando-o e agrupando-o com seus semelhantes. A

representação temática concentra-se nos assuntos do documento, obtidos a partir da análise de seu conteúdo. É um processo ligado aos conceitos da obra, sumarizados com o objetivo de possibilitar sua localização.

A representação da informação é um processo de suma importância na recuperação da informação, pois fornece condições para que as informações registradas nos documentos sejam localizadas em um sistema de informação. No contexto da Web, com a liberdade de produção de informações, características como autor, data de publicação, edição e outros dados extrínsecos ao documento podem não ser informados durante a publicação do documento, tornando difícil uma representação descritiva, ao passo que a representação temática, que pode identificar os assuntos, desde que feita de forma automática, é uma alternativa viável.

Com a produção e disponibilização quase instantânea de documentos na Web, processos como a indexação automática produzem resultados satisfatórios na representação da informação, tornando-se amplamente adotados por mecanismos de busca. No entanto, a qualidade destas representações pode ser questionável, resultando em poucos pontos de acesso a um documento.

Fujita (2012, p.13) afirma que a indexação e a recuperação da informação possuem uma estreita relação, uma relação de causa e efeito. Isto se dá, pois a indexação possui características que provocam efeitos sobre a recuperação. A especificidade é uma delas, pois a escolha de termos específicos, que serão traduzidos em descritores igualmente específicos para que posteriormente o usuário possa fazer busca por palavras chave específicas. Esta é apenas uma das decisões que devem ser tomadas durante a indexação e que refletem diretamente no resultado da recuperação.

Em outro extremo da Figura 1 está o **Usuário**, um elemento fundamental do processo de Recuperação da Informação, que tem em suas necessidades informacionais um objeto norteador de pesquisas na área da RI.

Dantas, Silva e Souza (2013) caracterizam os usuários como indivíduos que utilizam serviços de busca por informação, com o objetivo de preencher suas lacunas informacionais, agregando novos conhecimentos ou ampliando-os para constituir novos comportamentos e exercer suas atividades cotidianas na vida em sociedade.

Riecken (2006, p. 55) afirma que “o usuário e os conteúdos estão no centro das preocupações da CI e não as tecnologias”, evidenciado assim sua importância no âmbito dos

processos de busca e recuperação de informação. Para Ferneda (2012, p. 18), a recuperação da informação é um processo de produção de sentido por parte do usuário, o qual utiliza a informação para a construção do conhecimento. Em uma abordagem mais subjetiva, Morris (1994) explica que a informação é parcialmente construída pelo usuário durante este processo de produção de sentido, e só existe fora dele de maneira incompleta. Portanto, os SRI devem ser modelados considerando as necessidades informacionais e o comportamento do usuário.

Os usuários expressam suas necessidades informacionais utilizando uma **Expressão de Busca**, que em sua forma mais trivial é formada por uma ou mais palavras. A capacidade de elaboração de uma boa expressão de busca pode influenciar decisivamente no sucesso ou fracasso em obter as informações desejadas.

Ferneda (2012, p. 18) define uma expressão de busca como:

> […] o meio que o usuário utiliza para comunicar ao sistema sua necessidade informacional. É composta geralmente por termos que podem ser especificados em linguagem natural ou por meio de algum tipo de linguagem artificial dependo dos recursos do sistema. Para se obter sucesso neste processo os usuários necessitam de ter um mínimo de conhecimento do tema de interesse e do seu vocabulário do domínio.

A elaboração da expressão de busca é uma etapa importante no processo de recuperação de informação, pois é por meio dela que o usuário comunica ao sistema suas necessidades informacionais. A tarefa de construção da expressão de busca sofre influência de vários fatores, principalmente inerentes ao próprio usuário, como seu comportamento, sua capacidade de utilizar o sistema, seu domínio da área de conhecimento, entre outros. Desta forma, ao utilizar a linguagem natural para expressar suas necessidades, um usuário pode produzir uma consulta com um conjunto inadequado de termos de busca, o que fará com que o sistema não seja capaz de recuperar informações relevantes, mesmo que elas existam.

Buscando reduzir estas dificuldades, o modelo de Ferneda (2013) possui uma etapa denominada **Representação da Expressão de Busca**, que o autor define como uma série de recursos que auxiliem a tradução de uma necessidade de informação em uma expressão de busca adequada. Dentre estes recursos pode-se citar a eliminação de palavras de baixo poder semântico (*stopwords*), como artigos, preposições e conjunções. A padronização de caracteres para maiúsculos ou minúsculos, a correção ortográfica ou a conflação, que consiste na representação de dois ou mais termos em um único, por meio de processos como o *stemming* ou a lematização (do inglês *lemmatization*).

Este recurso de pré-processamento textual, que emprega técnicas de normalização linguística para possibilitar melhorias na compreensão da linguagem humana por parte dos algoritmos computacionais, pode apresentar benefícios significativos. Porém, eles podem não ser suficientes em determinadas situações, como a baixa quantidade de termos informados pelos usuários em razão da falta de conhecimento sobre o domínio pesquisado.

O método proposto nesta tese pode ser classificado como um recurso que auxilia o usuário nesta etapa do processo de Recuperação da Informação, na elaboração da expressão de busca e de sua representação. Por meio da apresentação de um elemento visual que contém, além da terminologia do domínio, sugestões de termos que podem ser agregados à consulta para torná-la mais apta a recuperar informações relevantes.

O elemento central do processo de recuperação de informação é a **Função de busca**, responsável por fazer a comparação entre as representações da expressão informada pelo usuário e cada documento do acervo, retornando a intersecção entre elas em formato de uma lista contendo os itens que supostamente contém as informações que o usuário procura (FERNEDA, 2012, p.19).

A função de busca tem papel crucial no processo de recuperação de informação, pois seu funcionamento impacta diretamente nos resultados apresentados ao usuário. Comparando as representações, da expressão de busca e dos documentos, alguns métodos se limitam a fazer uma comparação simples se os termos da expressão aparecem (ou não) no documento, resultando assim em uma lista com documentos que atendem aos critérios. Este tipo de método não possibilita uma ordenação dos resultados, o que não é desejável. Já outros métodos utilizam a atribuição de pesos aos termos e possibilitam assim uma comparação com o atendimento parcial dos critérios, possibilitando uma ordenação dos documentos mais relevantes frente a uma busca.

Na Figura 1 três elementos estão rodeados por uma linha tracejada, na representação de Ferneda (2012) este conjunto forma o **Modelo de Recuperação de Informação**, assim os métodos supracitados são definidos pelo modelo adotado. O autor ainda afirma que alguns deles foram criados há décadas, entretanto, suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e também nos mecanismos de busca da Web.

O último elemento é o **Resultado da busca**, que, em geral, é composto pelo conjunto de documentos que supostamente serão úteis para o usuário sanar sua necessidade de informação. É desejável que o resultado seja apresentado de forma a facilitar a obtenção da

informação desejada, uma alternativa bastante utilizada é a apresentação em formato de uma lista ordenada pelo grau de importância em relação à expressão de busca (FERNEDA, 2012, p.19).

Espera-se que essa listagem de registros recuperados possibilite ao usuário verificar se os documentos lhe são úteis em relação a sua necessidade de informação, pois alguns fatores podem tornar um documento integrante do resultado irrelevante, tais como como documentos obsoletos ou previamente recuperados e já apreciados por este indivíduo.

No contexto da Web ainda há outros fatores que influenciam nos resultados de uma busca, como adição artificial de resultados, posicionados ao topo da listagem, por meio do "patrocínio", ou seja, por meio do pagamento ao SRI alguns documentos serão apresentados nas primeiras posições dos resultados de busca. Há também a apresentação de documentos inacessíveis, comumente conhecidos como "links quebrados", que são resultados que apontam para documentos que por alguma razão não estão mais disponíveis.

A discussão sobre o processo de recuperação de informação, seu funcionamento e fatores que influenciam em seus resultados é bastante ampla, há incontáveis pesquisas com abordagens dedicadas a cada um dos elementos de forma individual ou agregando duas ou mais etapas, no entanto, algo que impacta diretamente em qualquer sistema de recuperação de informação é o modelo que este utiliza, por essa razão faz-se na sequência uma breve explanação de cada um dos modelos clássicos supracitados.

2.2 Os modelos clássicos

Apresentar uma lista de documentos que contenham informações úteis como resposta a busca de um usuário é um dos objetivos primários de um sistema de recuperação de informação. Baeza-Yates e Ribeiro-Neto (2013) explicam que os sistemas utilizam algoritmos com procedimentos e características previamente definidas para julgar a relevância e indicar uma ordenação utilizando um conjunto de critérios estabelecidos. Estes procedimentos e regras pré-definidos nos algoritmos formam os modelos de RI.

Existem modelos baseados apenas no texto de cada documento, que o utilizam para ranqueá-lo em relação à consulta, são chamados de modelos não estruturados. Os modelos estruturados consideram também a estrutura do documento, ponderando a ocorrência de termos nos títulos, palavras-chave e corpo do documento. Há ainda alguns modelos que foram criados especificamente para o ambiente Web, que utilizam por exemplo as ligações entre páginas para

fornecer indícios de relevância de cada registro, produzindo uma listagem mais adequada (BAEZA-YATES; RIBEIRO-NETO, 2013).

Os modelos de recuperação de informação vêm sendo aprimorados há anos, mesmo antes do advento dos computadores e dos recursos tecnológicos atuais. Alguns como o Modelo Booleano, o Modelo Vetorial e o Probabilístico são considerados "os modelos clássicos de recuperação de informação" e até hoje possuem conceitos úteis para as pesquisas na área da RI (SILVA; SANTOS; FERNEDA, 2013, p. 29).

O modelo de recuperação de informação booleano tem seus fundamentos na Teoria dos Conjuntos e na Álgebra Booleana (ou álgebra de Boole, em homenagem ao matemático inglês George Boole). É considerado um dos modelos mais simples e intuitivo. Seu funcionamento é baseado na existência ou não dos termos da expressão de busca em cada documento, utilizando o sistema binário o método funciona atribuindo os valores "1" e "0" para os termos presentes e ausentes respectivamente.

A expressão de busca neste modelo é composta por um ou mais termos que podem ser ligados pelos conectivos lógicos *AND*, *OR* e *NOT*, formando assim uma expressão booleana que submetida a um conjunto de documentos elenca quais satisfazem as restrições determinadas (BAPTISTA, 2019, p.42). O alcance das consultas no modelo booleano baseia-se na utilização dos operadores, sendo o operador "*OR*" utilizado para ampliação e o operador "*AND*" para restrição.

Uma das desvantagens deste modelo é que devido ao seu critério binário, os documentos são classificados apenas como relevantes ou irrelevantes para a busca, impossibilitando alguma forma de ordenação dos resultados (SOUZA, 2006). Essa característica do modelo booleano faz com que a tarefa de em encontrar informações úteis seja mais complicada, pois a listagem dos documentos não é apresentada de acordo com seu grau de utilidade.

Kuramoto (2002) ainda aponta outro problema significativo do modelo, a dificuldade em formular expressões de busca utilizando adequadamente os operadores booleanos, visto que essa é uma tarefa pouco trivial para usuários leigos. A falta de domínio da lógica booleana implica em expressões simples e/ou genéricas que podem resultam em grandes volumes de resultados. Ainda nesse sentindo o autor afirma que há pouco controle sobre a quantidade de documentos recuperados, sendo necessárias subsequentes reformulações da expressão de busca para atingir uma quantidade de resultados aceitável.

O modelo do espaço vetorial ou simplesmente modelo vetorial foi proposto por Salton (1971) e tinha como meta superar as limitações do modelo booleano. Baeza-Yates e Ribeiro-Neto (2013) relatam que o método propõe um quadro onde casamentos parciais são possíveis por meio da atribuição de pesos não binários aos termos das consultas e dos documentos, desta forma é possível efetuar o cálculo do grau de similaridade entre cada documento da base e a expressão de busca, originando uma listagem de documentos decrescente de acordo com o grau de similaridade.

O nome do modelo origina-se da representação dos documentos e das consultas por vetores em um sistema de coordenadas cartesianas. Cada um dos termos recebe um peso que indica sua relevância. Ferneda (2012, p.31) explica que estes pesos devem ser normalizados para assumir valores entre 0 e 1, sendo que os pesos próximos de 1 indicam uma maior importância e os pesos próximos de 0 uma menor importância.

O funcionamento do modelo consiste em determinar a similaridade entre cada documento do acervo e a expressão de busca utilizando cálculo do produto escalar entre os vetores que os representam. Esta dinâmica de funcionamento torna este modelo uma boa opção para lidar com grandes coleções, pois estes tipos de cômputos matemáticos são feitos com facilidade pelos computadores modernos.

Para determinar os pesos dos termos de indexação no modelo vetorial são utilizadas várias estratégias, visando atribuir valores mais altos aos termos que supostamente possuem maior capacidade de distinção dentro dos documentos. Ferneda (2013, p.33) explica que “os melhores termos de indexação (que apresentarão maiores peso) são aqueles que ocorrem com uma grande frequência em poucos documentos”. Para determinar estes pesos em cada um dos termos são utilizadas as medidas TF (*Term Frequency*), que representa a frequência do termo e IDF (*Inverse Document Frequency*) representa a frequência inversa do termo.

Salton e McGill (1983. p. 204-207), definem TF como o número de vezes que um determinado termo aparece no texto de um documento, e, IDF como a divisão do número de documentos no corpus pelo número de documentos que contém o termo. A medida IDF consiste em uma estatística global do acervo utilizada para distinção da utilidade de um termo em relação à coleção.

Um dos conceitos principais do modelo vetorial é esse cômputo de pesos para os termos de indexação e sua consequente vantagem de ordenação de resultados de acordo com o grau de pertinência em relação à expressão de busca. No contexto do ambiente Web, onde o acervo é

imensurável a ordenação dos resultados de busca e o posicionamento dos documentos mais relevantes nas primeiras posições é uma característica fundamental.

Acredita-se que um mecanismo de busca Web deve utilizar este princípio para que seja capaz suprir as necessidades informacionais de seus usuários satisfatoriamente, sendo que o método proposto neste estudo necessita de um modelo capaz de ordenar seus resultados de busca pela relevância.

O terceiro e último modelo clássico é o probabilístico, que teve sua primeira menção em 1960 por Maron e Kuhns e, posteriormente, em 1976 por Robertson e Sparck Jones, que descreveram uma variação do modelo para recuperação de informação utilizando o arcabouço probabilístico (BAEZA-YATES; RIBEIRO-NETO, 2013; FERNEDA, 2012).

O modelo utiliza probabilidades para avaliar o grau de relevância entre a expressão de busca e o corpus documental. Seu funcionamento baseia-se no princípio de que para cada consulta de um usuário existe um conjunto de documentos composto unicamente por documentos relevantes, denominado resposta ideal. Dado o fato de que estes documentos são representados por termos, que foram atribuídos na etapa de indexação, há uma expressão de busca capaz de descreve-los e recuperar essa resposta ideal.

O usuário não sabe exatamente quais são os termos necessários para obter a resposta ideia, sendo assim ele formula uma expressão de busca e a submete ao sistema. Ao retornar o conjunto de documentos correspondente o sistema permite que o usuário classifique quais documentos são ou não relevantes para sua necessidade de informação. O sistema utiliza esse *feedback* para fazer uma nova busca e recuperar um novo conjunto de documentos com maior probabilidade de serem úteis. Este processo pode ser repetido até que o usuário julgue que encontrou sua resposta ideal.

A possibilidade de atribuição de relevância por parte do usuário é uma característica significativa do modelo probabilístico, que inspirou a técnica de *relevance feedback*. Pottker (2017) relata que por meio de sucessivas interações com o sistema, usuários podem alcançar de forma gradativa, resultados mais precisos e relevantes para suas necessidades de informação.

Uma limitação do modelo probabilístico é que ele não é capaz de auxiliar o usuário na formulação da consulta inicial, na primeira iteração. Souza (2006) afirma que ele pode utilizar na tentativa inicial técnicas de outros modelos, como o vetorial por exemplo, e que depois de sucessivas interações e do *feedback* do usuário os resultados de busca tornam-se cada vez mais

próximos do conjunto ideal, sendo este o maior destaque do modelo, considerar o usuário como fator decisivo para alcançar os resultados ideais.

O princípio da utilização do *feedback* do usuário para o aprimoramento dos resultados de busca é a característica mais marcante deste método. Acredita-se que este é o caminho mais promissor para o aprimoramento dos mecanismos de busca atuais, incluir o usuário e auxiliá-lo a descrever de forma mais eficiente sua necessidade informacional.

O método proposto explora essa ideia, de que uma maior interação com o sistema pode fornecer dados valiosos no aprimoramento da expressão de busca e consequentemente nos resultados da busca. Tanto o processo de contextualização quanto a expansão de consultas descritas neste estudo baseiam-se na comunicação do usuário com o SRI.

2.3 A Recuperação de Informação no ambiente Web

Baeza-Yates e Ribeiro-Neto (2013, p.3) afirmam que “apesar de sua maturidade, até recentemente a RI era vista como uma área de interesse limitado a bibliotecários e especialistas em informação”. No entanto, no início dos anos 90, com o surgimento da *World Wide Web* e da troca de informações em âmbito global essa situação mudou completamente (BAEZA-YATES; RIBEIRO-NETO, p.4).

O surgimento da Internet e do ambiente Web possibilitou a criação de um repositório universal da cultura e do conhecimento humano, pois qualquer indivíduo pode criar e publicar documentos, ocasionando uma expansão do volume de informações sem precedentes. Uma consequência imediata foi que encontrar informações úteis na Web tornou-se uma tarefa mais complexa, fazendo com que a área de RI e suas pesquisas ganhassem notoriedade no meio acadêmico e empresarial.

No entanto, o processo de recuperação de informação no ambiente Web possui particularidades e desafios únicos, devido a fatores como o volume de informação disponível, a heterogeneidade de idiomas e linguagens, a democratização de seu acesso além da importância que a Internet tomou dos dias atuais.

Baeza-Yates e Ribeiro-Neto (2013, p.4) afirmam que o objetivo principal de um SRI é “recuperar todos os documentos que são relevantes à necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes”. Neste sentido, o objetivo de um mecanismo de busca Web não é diferente, sua função é servir de

ferramenta intermediária entre um indivíduo com uma necessidade informacional e os documentos que contenham as informações que satisfaçam essa necessidade.

A utilização da Internet como meio de obter informações vem se tornando cada vez mais comum, de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), no ano de 2019, a Internet era utilizada por 82,7% dos domicílios brasileiros, sendo o celular o equipamento mais utilizado para o acesso (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2021).

A mesma pesquisa aponta que dentre as tarefas desempenhadas pelas pessoas na Internet destacam-se a comunicação em mensageiros eletrônicos, indicada por 95,7% das pessoas, o consumo de entretenimento audiovisual, com 88,4% e o envio e recebimento de e-mails com 61,5%.

Outra pesquisa, com o foco em educação, observou que dentre os estudantes de escolas localizadas em áreas urbanas, 79% utilizavam a Internet para buscar notícias, 88% para buscar informações para fazer algo que não sabiam ou sentiam dificuldades para fazer e 93% para fazer pesquisas relacionadas à trabalhos escolares (NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR, 2020).

Para se fazer pesquisas na Internet geralmente utiliza-se algum mecanismo de busca, cujo representante mais conhecido e mais utilizado é o Google Busca[1], que de acordo com a Statcounter, em maio de 2021, o serviço da Google respondia por 92,2% de todas as buscas feitas na internet (STATCOUNTER GLOBAL STATS, 2021).

No entanto, apesar de sua importância no processo de busca da informação, os buscadores existentes, baseados em caixas de texto para construção da expressão de busca utilizando palavras-chave, apresentam falhas no atendimento das expectativas dos usuários. Baeza-Yates e Ribeiro-Neto (2013, p.410-412) entendem que os principais desafios da recuperação de informação na Web podem ser divididos em duas grandes categorias: os problemas relacionados aos dados e os relacionados com os usuários e suas interações com os dados.

Dentre os desafios relacionados aos dados estão:

[1] O Google Busca é um serviço da empresa Google que permite fazer pesquisas na internet sobre quaisquer conteúdos. Disponível em: https://www.google.com.br/. Acesso em: 26 jun. 2021.

- alto percentual de dados voláteis: dados podem ser adicionados ou removidos facilmente, ocasionando por exemplo links suspensos (quebrados);
- grande volume de dados: a escala de crescimento dos dados na Web se mostra um desafio difícil de lidar;
- dados redundantes: devido a sua natureza democrática, a Internet possibilita que cópias de conteúdos sejam publicadas aumentando o volume de dados, porém com informações repetidas;
- baixa qualidade dos dados, que é justificada por registros imprecisos, obsoletos ou inválidos. Há ainda a ocorrência de informações erradas produzidas de forma inocente ou maliciosa;
- dados heterogêneos e não estruturados, há uma grande quantidade de formatos e linguagens que os dados podem ser disponibilizados. Além disso em uma parte significativa deles não há uma estrutura que permita sua interpretação por algoritmos e agentes inteligentes.

Os autores explicam que desafios como redundância e a má qualidade dos dados podem ser solucionados com o aprimoramento dos algoritmos. No entanto, alguns desafios simplesmente não podem ser resolvidos (como a diversidade de linguagens e dialetos utilizados pelas pessoas), pois são intrínsecos à natureza humana, tornando-se assim desafios permanentes na recuperação da informação.

A segunda classe de desafios está relacionada a interação do usuário com o SRI e é composta basicamente por dois aspectos:

- construção da expressão de consulta: frequentemente a elaboração de uma expressão capaz de refletir as necessidades informacionais do indivíduo pode não ser simples;
- interpretação dos resultados: mesmo considerando que o usuário seja capaz de expressar perfeitamente suas necessidades em uma consulta, o resultado pode conter milhares de documentos ou não recuperar nada. No caso de grandes quantidades de documentos, como ranquear adequadamente? Quais documentos são relevantes? Como navegar por eles?

Estes problemas são desafios específicos do ambiente Web, porém são antes disso dificuldades que podem ter soluções nas pesquisas da área de RI, sendo assim úteis como insumos para o desenvolvimento de estudos como esta tese.

Fachin (2009) afirma que um dos principais problemas dos sistemas de recuperação de informação é recuperar somente documentos importantes para o usuário, ou seja, documentos relevantes. No entanto, muitos sistemas possuem apenas relevância parcial. Para superar essas limitações diversas abordagens são possíveis, sendo que o emprego de tesauros ou ontologias no tratamento padronizado e representação da informação implica em melhorias na correspondência entre a expressão de busca e as representações dos documentos, proporcionando uma recuperação de informação mais precisa e útil ao usuário.

Desta forma, as ontologias são vistas como soluções para a recuperação relevante da informação, pois são correlatas diretas dos tesauros, e podem ser empregadas nas etapas de classificação, indexação e tratamento linguístico/semântico de cada termo a ser incorporado na representação das informações (FACHIN, 2009, p. 268).

Diferentes pesquisas foram desenvolvidas buscando melhorar a recuperação de informação utilizando tecnologias da Web Semântica (VALLET; FERNÁNDEZ; CASTELLS, 2005; CASTELLS; FERNÁNDEZ; VALLET, 2006; BHAGDEV *et al.*, 2008). Estes trabalhos se enquadram no campo de pesquisa conhecido como busca semântica, pois referem-se a sistemas que usam tecnologias da Web Semântica para melhorar diferentes partes do processo de RI (PABÓN; GONZÁLEZ, 2014, p. 53).

Para Mangold (2007), a busca semântica pode ser definida como um processo de recuperação de documentos que tira proveito do conhecimento de um domínio formalizado por meio de uma ontologia. Na perspectiva de Wei, Barnaghi e Bargiela (2008), a busca semântica estende o escopo dos paradigmas tradicionais de recuperação de informação, possibilitando a recuperação de entidades e conhecimentos por meio da representação do significado das palavras obtido a partir de sua formalização utilizando ontologias.

Fernández *et al.* (2011) apresentam uma abordagem para recuperação conceitual de informação baseada em ontologias. Os autores buscam superar as limitações de expressões de busca compostas apenas por palavras-chave utilizando as ontologias para capturar e compreender o significado, o conceito das necessidades dos usuários. Por meio da proposição de um modelo de recuperação de informações são executados experimentos que demonstram resultados positivos.

O estudo de Rinaldi (2009) consiste em uma métrica para medir a relação semântica entre palavras utilizando ontologias. Explorando propriedades linguísticas a abordagem classifica os documentos obtidos em resposta a uma busca, ordenando-os de acordo com os interesses do usuário. Experimentos em uma base de testes mostraram melhorias efetivas na recuperação de informação.

Vijayarajan *et al.* (2016) propõem uma estrutura genérica para recuperação de imagens na Web baseada em ontologias. O estudo utiliza uma ontologia construída com descrições de imagens para fornecer ao usuário um modelo de dados que atua como um dicionário para refinar suas palavras-chave e reduzir a lacuna semântica entre os resultados recuperados e os esperados.

O denominador em comum entre estes estudos é o emprego de ontologias no processo de recuperação de informação, com objetivo de superar alguns dos desafios e dificuldades da busca de informação na Web, tal como a dificuldade em elaborar consultas que representem precisamente as necessidades de informação em mecanismos de busca. Sendo assim o próximo capítulo discorre sobre as ontologias, suas definições, características e aplicações.

3 Ontologias

Este capítulo apresenta uma discussão sobre as ontologias, artefatos fundamentais para esta pesquisa em razão de função primordial: a representação do conhecimento com vistas a sua organização, acesso e apropriação. Destaca-se inicialmente a origem do termo no âmbito Filosófico, suas definições nas áreas da ciência da informação e da computação e na sequência suas características e componentes. São apresentados ainda alguns detalhes técnicos da linguagem OWL que são essenciais à compreensão do método proposto.

Um dos princípios fundamentais da comunicação está relacionado ao nível de intelecção na emissão e na recepção da mensagem, ou seja, não será possível estabelecer uma comunicação sem algum grau de compreensão mútua. (MOREIRA, 2010, p.18)

As condições para a comunicação efetiva da informação e do conhecimento formam, na definição clássica de Saracevic (1996, p. 47, grifo nosso) uma das preocupações centrais da Ciência da Informação, que pode ser entendida como

> um campo dedicado às questões científicas e à prática profissional **voltadas para os problemas de efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual** do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais

Para que uma conceituação possa ser comunicada entre pessoas e também entre algoritmos ela deve ser expressa em termos de um artefato concreto representado em uma linguagem. As ontologias, fundamentadas na modelagem de domínios, permitem a explicitação dos compromissos ontológicos que representam um domínio, agregando fidelidade, consistência e clareza nas representações (CAMPOS, 2013, p. 133).

Compromisso ontológico é um termo formulado por Willard Van Orman Quine, um filósofo e matemático estadunidense que se dedicou ao estudo da filosofia analítica (WILLARD VAN ORMAN QUINE, 2021). Quine propõe uma expressão formal desse compromisso pelo uso de quantificadores lógicos, para o autor, algo existe em uma teoria científica, se e somente se, puder ser formalizado pela lógica de predicados. Sendo assim o compromisso é fundamental em uma teoria científica, para seu uso posterior na conceituação do mundo em ontologias (BAX; COELHO, 2012, p. 2).

Nodine e Fowler (2005) definem compromisso ontológico como um acordo firmado por uma comunidade sobre o significado que se estabelece nas representações de conceitos, tanto do ponto de vista da compreensão pelo homem quanto do tratamento pela máquina. Isso implica em definir o vocabulário de uma forma que venha a minimizar ambiguidades, para que seu uso possa ser partilhado na representação e recuperação do conhecimento.

Guarino, Carrara e Giaretta (1994) situam o papel do compromisso ontológico como o de um elemento fomentador da precisão entre a conceituação e a representação de uma visão de mundo, esta última, essencialmente imprecisa em algum grau em relação ao significado pretendido pelo homem. Essa imprecisão se dá devido ao fato de que as conceituações são entidades abstratas, que existem na mente de pessoas ou grupo de pessoas de uma comunidade (GUIZZARDI, 2007).

A dificuldade em representar conceitos decorre, ao menos em parte, das limitações das linguagens, que não conseguem ser suficientemente expressivas para representar formalmente a riqueza semântica da conceituação presente na mente humana (CAMPOS; CAMPOS; MEDEIROS, 2011). No entanto, mesmo considerando essa limitação, a aproximação dos modelos conceituais com a realidade, ainda se mostra essencial à comunicação das informações, e a capacidade de dotar algoritmos com uma mínima compreensão de um modelo de mundo com seus conceitos, relações, axiomas e indivíduos.

As ontologias, objetos que incorporam esse compromisso ontológico vem se tornando foco de interesse em várias abordagens que buscam explorar estes conhecimentos e utilizá-los de variadas formas para aprimorar a construção de representações das informações.

3.1 Breve histórico e conceituação do termo ontologia

A primeira menção do termo ontologia é atribuída ao filósofo e pedagogo Jacob Lorhard (*Jacobus Lorhardus*) (1561-1609) em sua obra *Ogdoas Scholastica*, de 1606. Mais tarde, em 1730 com a publicação da obra *Philosophia prima sive Ontologia*, de Christian Wolff (1679-1754), o termo ontologia tomou visibilidade nos círculos filosóficos, sendo considerado sinônimo de *metaphysica generalis* parte da metafísica que analisa as características do ser em geral (FERNEDA, 2013, p.44).

Em seu sentido original, a palavra ontologia quando empregada na área da filosofia faz referência ao estudo do ser enquanto ser, da natureza da existência. Sua etimologia é definida por Castro (2008, p.7):

> Ela é o resultado da junção de dois termos gregos *onta* (entes) e *logos* (teoria, discurso, palavra). Ao pé da letra, ontologia significa, portanto, teoria dos entes. “Ente” está aí representando todas as coisas sobre as quais se pode dizer que são – ou que a ontologia é a teoria do ser enquanto tal.

Em uma busca pela palavra ontologia nos dicionários da língua portuguesa Houaiss e Villar (2009), Priberam (2021), Michaelis... (2021), Dicio (2021) e Aulete Digital (2014) é possível observar um consenso nas definições da filosofia e da medicina, sendo que apenas Houaiss e Villar (2009) não apresentam uma definição do campo da informática

> **on·to·lo·gi·a**
>
> (onto- + -logia)
>
> *substantivo feminino*
>
> 1. [Filosofia] parte da filosofia que tem por objeto o estudo das propriedades mais gerais do ser, apartada da infinidade de determinações que, ao qualificá-lo particularmente, ocultam sua natureza plena e integral; metafísica ontológica
> 2. [Medicina] Doutrina do século XIX que estuda o ser da patologia, tendo em conta a maneira como ela se original, como se a enfermidade existisse em conformidade a um tipo bem definido, a uma essência.
> 3. [Informática] Técnica de organização de informação, composta por um conjunto estruturado de termos, conceitos e suas relações que representam um determinado conhecimento, buscando compartilhar e reutilizar essas informações. (DICIO, 2021, Não paginado).

O dicionário Aulete Digital (2014) é o único que frisa na definição do campo da informática que as ontologias conceituam o conhecimento de forma explícita e formal, ou seja, processável por máquinas e compartilhável.

Uma definição recorrente encontrada com frequência em textos que tratam sobre Web Semântica é a de Gruber (1995, p. 908, tradução nossa), que define ontologia como “uma especificação formal e explícita de uma conceitualização compartilhada”, onde, a conceitualização representa um modelo abstrato de algum fenômeno que identifica os conceitos relevantes para si mesmo; explicita, pois, os elementos e suas restrições são claramente definidas, formal na medida em que se deve ser passível de processamento automático (por máquinas) e compartilhada, por ser composta por conhecimento consensual, aceito por um grupo de pessoas.

Outra definição clássica, a de Guarino (1998, p.4, tradução nossa) explica que no contexto da inteligência artificial uma ontologia

> refere-se a um artefato de engenharia, constituído por um vocabulário específico usado para descrever uma certa realidade, mais um conjunto de suposições explícitas sobre o significado pretendido das palavras do vocabulário. Esse conjunto de suposições geralmente tem a forma de uma teoria lógica de primeira ordem, na qual as palavras do vocabulário aparecem como nomes de predicado unários ou binários, respectivamente chamados conceitos e relações. No caso mais simples, uma ontologia descreve uma hierarquia de conceitos relacionados por relacionamentos de subsunção; em casos mais sofisticados, axiomas adequados são adicionados a fim de expressar outras relações entre concepções e restringir sua interpretação.

Em ambas as definições é possível identificar características marcantes das ontologias, a existência de uma estrutura hierárquica composta por conceitos e relações e a possibilidade de ser compreendida por seres humanos e máquinas computacionais.

Outro aspecto importante das as ontologias é que elas são, antes de tudo, vocabulários de representação geralmente especializados em algum domínio. Elas fornecem termos em potencial para descrever o conhecimento sobre alguma área específica do conhecimento, sendo assim são particularmente úteis no entendimento da linguagem natural (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999).

Jasper e Uschold (1999, p.11-2, tradução nossa, grifo nosso) ressaltam a necessidade de explicitação dos relacionamentos entre os termos (conceitos) pertencentes a uma ontologia:

> Uma ontologia pode possuir uma variedade de formas, mas necessariamente incluirá um vocabulário de termos, e alguma especificação de seus significados. Isto inclui **definições e uma indicação de como os conceitos estão inter-relacionados, o que impõem uma estrutura no domínio e restringe possíveis interpretações de termos**.

O trecho destacado evidencia a importância de uma construção criteriosa das ontologias, para que seja possível além da compreensão de cada conceito uma especificação sólida dos relacionamentos entre eles, para que as representações estejam livres de ambiguidades e interpretações dúbias.

Um aspecto central das ontologias é sua relação com a organização do conhecimento, explicitada por autores como Mizoguchi (2004) que a definem como um conjunto de conceitos fundamentais e suas relações, que captam como as pessoas entendem ou interpretam um domínio do conhecimento, permitindo a representação desse entendimento de maneira formal, compreensível por humanos e computadores.

Moreira (2010, p.28) ressalta que uma das funções das das ontologias é fornecer sistemas de categorização que permitam ao homem organizar formalmente sua realidade,

preocupando-se essencialmente com a identificação das características comuns a todos os seres, e, por meio da observação, possibilitando o conhecimento do mundo físico, do raciocínio e produzindo uma estrutura de abstrações.

Nas definições de Mizoguchi (2004) e Moreira (2010) pode-se observar um aspecto importante das ontologias, a capacidade de armazenar o entendimento de um grupo de indivíduos sobre determinado domínio do conhecimento, que resulta na materialização da cognição humana em formato disponível para acesso de outras pessoas e algoritmos computacionais.

Desta forma compreende-se como as ontologias são instrumentos úteis aos processos de organização e transmissão do conhecimento registrado, sendo que que na Ciência da Informação elas são classificadas como sistemas de organização do conhecimento (SOC).

Para Brascher e Café (2008, p. 8), os SOC são "sistemas conceituais que representam determinado domínio por meio da sistematização dos conceitos e das relações semânticas que se estabelecem entre eles", definição que guarda semelhanças significativas com as definições de ontologias supracitadas.

Vickery (2011) complementa que as ontologias fazem parte de um grupo de SOC pertencentes a Era da Web Semântica, pois diferenciam-se dos demais ao serem projetados para uso de agentes inteligentes, ou de forma mais genérica, para uso em sistemas computacionais.

Alguns autores como Broughton *et al*. (2004, p.143, tradução nossa) afirmam que "os SOC são ferramentas semânticas constituídas por palavras, conceitos e relações semânticas", no entanto, assim como Brascher e Carlan (2010, p. 152) prefere-se a adoção do vocábulo "termos" ao invés de "palavras".

Essa diferenciação entre "termos" e "palavras" faz-se necessária pois os vocábulos exprimem ideias distintas. O Tesauro da Ciência da Informação, que tem como base o Tesauro de Ciência da Informação, Tecnologia e Biblioteconomia (*ASIS&T Thesaurus of Information Science, Technology and Librarianship*) define os termos como "palavras ou frases usadas para denotar conceitos" (REDMOND-NEAL; HLAVA, 2005, p.128, tradução nossa), enquanto Cervantes (2020, p.9) complementa que eles são considerados "unidades de base da Terminologia".

Desconsiderando a complexidade na definição de “palavra” considera-se que ela pode ser entendida com “uma unidade lexical, uma sequência fônica se associa de modo relativamente estável a um significado ou conjunto de significados” (BASILIO, 2016, p. 10).

Cunha (2013) explica que uma palavra é constituída de elementos materiais que possui um sentido a que se presta, enquanto os vocábulos referem-se aos elementos materiais que o constituem, sendo assim, “na linguagem corrente os dois termos, palavra e vocábulo se equivalem” (PEREIRA, 2013, p. 7).

3.2 Características e componentes

De forma geral a construção de uma ontologia pode ser pensada como uma união de elementos que formam uma estrutura complexa. Classes e subclasses são arranjadas em uma estrutura hierárquica que é complementada por propriedades descritivas e relacionais, além de regras, axiomas e instâncias, que são outros elementos que compõem as ontologias.

Uma das funções de uma ontologia é reunir os conceitos utilizados em uma determinada área do conhecimento, padronizando seus significados e apresentando definições comumente aceitas por especialistas da área. Segundo Noy e McGuinness (2001) seu desenvolvimento consiste essencialmente na estruturação das classes em uma hierarquia taxonômica, na definição das propriedades ou atributos das classes e a na criação das regras que descrevem os valores permitidos para estes atributos.

O Quadro 1 apresenta uma definição dos elementos que compõem as ontologias segundo a perspectiva de Noy e McGuinness (2001) e de Ramalho (2010), permitindo assim uma complementação das definições.

Quadro 1 – Definição dos componentes de uma ontologia

	Noy e McGuinness (2001)	Ramalho (2010)
Classes	As classes representam os conceitos do domínio, são muitas vezes o foco das ontologias. Dentro de uma classificação hierárquica podem se subdividir originando as subclasses, que por sua vez herdam as propriedades da classe a que pertencem.	As classes e subclasses de uma ontologia agrupam um conjunto de elementos, "coisas", do "mundo real", que são representadas e categorizadas de acordo com suas similaridades, levando-se em consideração um domínio concreto. Os elementos podem representar coisas físicas ou conceituais, desde objetos inanimados até teorias científicas ou correntes teóricas;
Propriedades descritivas	São as várias características e atributos que descrevem cada conceito, as propriedades da classe.	Descrevem as características, adjetivos e/ou qualidades das classes;
Propriedades relacionais	São as restrições das propriedades, que podem descrever os tipos de valores, valores permitidos, números de valores (cardinalidade) e outras características que os valores podem ter.	Especificam os relacionamentos entre classes pertencentes ou não a uma mesma hierarquia, descrevendo e rotulando os tipos de relações existentes no domínio representado;
Regras e axiomas		Enunciados lógicos que possibilitam impor condições como tipos de valores aceitos, descrevendo formalmente as regras da ontologia e possibilitando a realização de inferências automáticas a partir de informações que não necessariamente foram explicitadas no domínio, mas que podem estar implícitas na estrutura da ontologia;
Instâncias	Representam os elementos de uma ontologia, ou seja, são as ocorrências dos conceitos e relações que foram estabelecidas pela ontologia. Uma instância é um conceito que pertence a uma classe e que possui determinados valores de atributos.	Indicam os valores das classes e subclasses, constituindo uma representação de objetos ou indivíduos pertencentes ao domínio modelado, de acordo com as características das classes, relacionamentos e restrições definidas;
Valores		Atribuem valores concretos às propriedades descritivas, indicando os formatos e tipos de valores aceitos em cada classe.

Fonte: Elaborado pelo autor

Observa-se que Noy e McGuinness (2001) não apresentam definições específicas para as regras, axiomas e valores, por esta razão as lacunas no quadro para as respectivas conceituações não estão preenchidas. Nos demais elementos: classes, propriedades e instâncias podem-se notar uma proximidade entre as definições de Noy e McGuinness (2001) e Ramalho (2010) indicando um consenso entre os autores.

Para Ferneda (2013, p.62) as ontologias são originadas pela

> […] união de peças que formam uma estrutura completa. Classes e subclasses definem um "esqueleto" na forma de uma hierarquia […], complementada por propriedades descritivas, propriedades relacionais, regras e axiomas. A sua

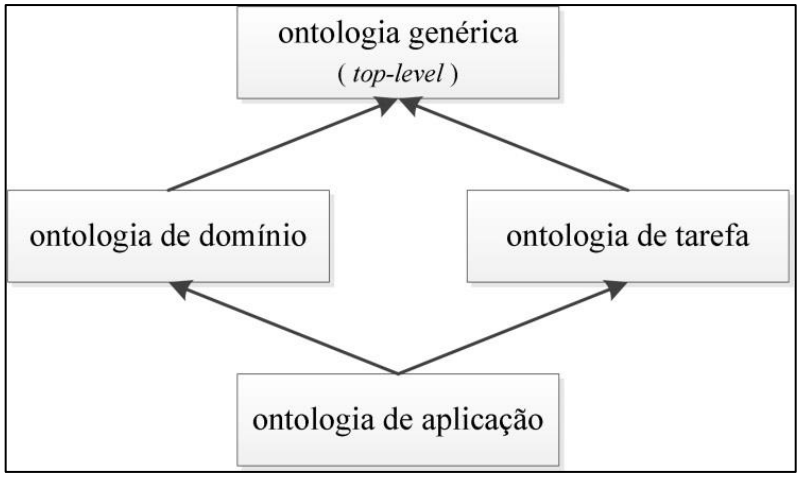
> abrangência (domínio) deve ser previamente definida e estabelece uma área do conhecimento ou uma parte do mundo que se pretende tratar.

A questão da abrangência das ontologias depende no nível de generalidade e do grau de formalismo ao qual elas são expressas, originando tipos de artefatos destinados a atender a propósitos distintos.

De acordo com Guarino (1998, p. 9), existem quatro tipos de ontologias: genéricas, de domínio, de tarefa e de aplicação, onde as ontologias de domínio ou de tarefa são especializações dos termos de uma ontologia genérica, ao passo que uma ontologia de aplicação deve ser composta por especializações dos termos do domínio ou tarefa correspondentes.

A Figura 2 apresenta uma representação gráfica destes tipos e seus níveis generalidade, sendo que as setas indicam os relacionamentos de especialização e dependência eles.

Figura 2 – Tipos de ontologias de acordo com o grau de generalidade



Fonte: Adaptado de Guarino (1998, p.9)

Ainda de acordo com Guarino (1998, p.9-10) cada nível pode ser descrito da seguinte forma:

- ontologias genéricas (*top-level ontologies*): descrevem conceitos muito gerais como espaço, tempo, matéria, objeto, evento, ação, etc., independentes de um problema ou domínio particular;
- ontologias de domínio e ontologias de tarefa: descrevem, respectivamente, o vocabulário relacionado a um domínio como medicina, automóveis, etc, ou uma tarefa ou atividade como diagnóstico, vendas, etc, especializando os termos introduzidos no nível superior;

- ontologias de aplicação: descrevem conceitos dependentes de ambos, um determinado domínio e tarefa particular. Estes conceitos frequentemente correspondem aos papéis desempenhados por entidades do domínio ao executar uma determinada atividade, como um guia para reposição dos componentes de um modelo de automóvel.

Além dos níveis de generalidade há ainda uma distinção entre o grau de formalismo do vocabulário utilizado na construção das ontologias. Uschold e Gruninger (1996, p. 97, tradução nossa) explicam que elas podem ser separadas em quatro categorias:

- altamente informais: expressas livremente em linguagem natural;
- semi-informais: expressas em linguagem natural de forma restrita e estruturada;
- semi-formais: expressas em uma linguagem artificial definida formalmente, como por exemplo Ontolíngua;
- rigorosamente formais: expressas por meio de termos meticulosamente definidos utilizando semântica forma, teoremas e provas de propriedades como consistência e completude.

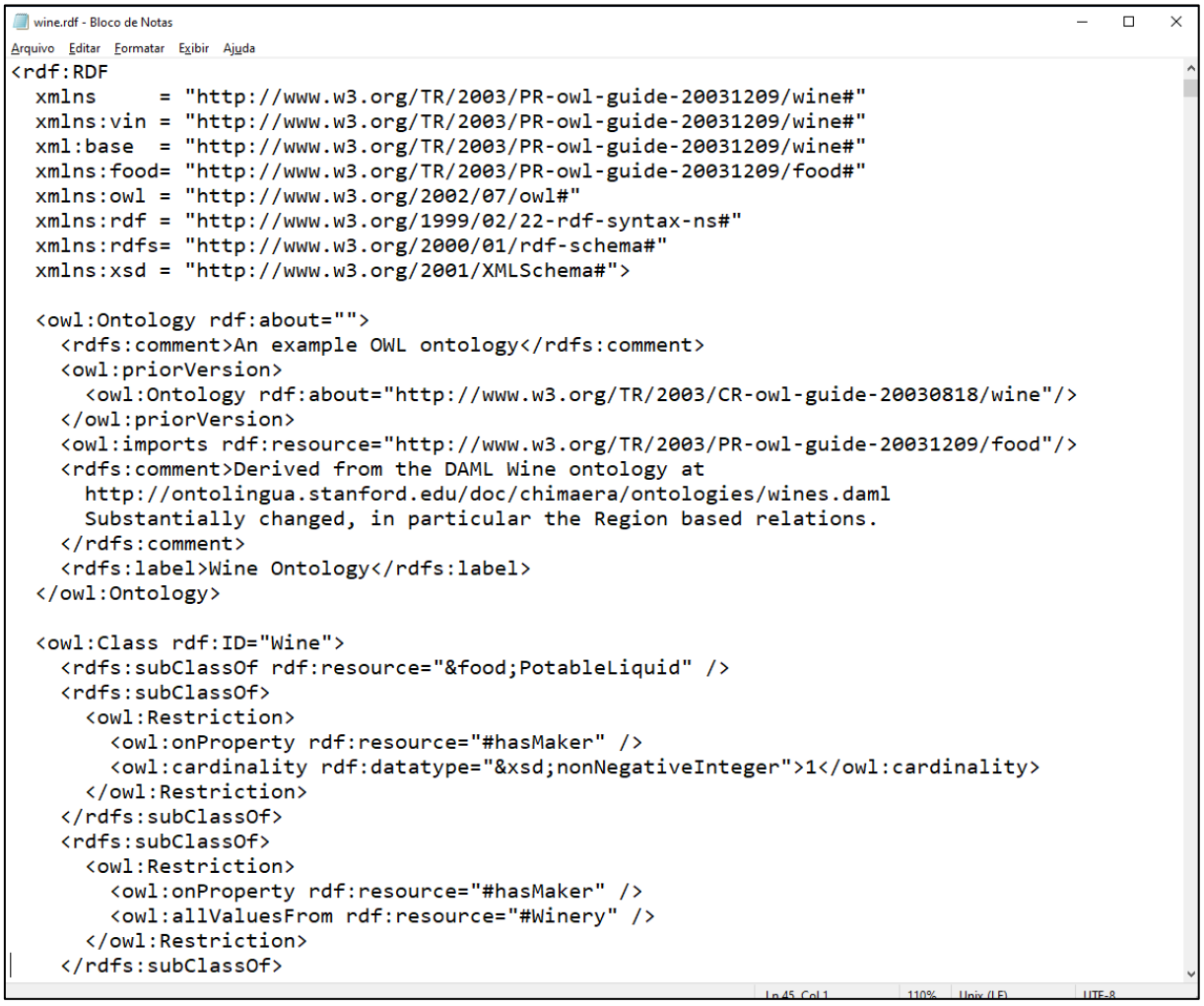
Para que sejam efetivamente úteis como instrumentos de organização do conhecimento as ontologias devem ser no mínimo semi-formais, empregando em sua construção linguagens artificiais que possuem capacidade de expressar as classes, propriedades e relacionamentos de forma compreensível por humanos e máquinas e possibilitando assim a solução de problemas de comunicação e interoperabilidade no que tange a heterogeneidade semântica (FARINELLI; ALMEIDA, 2014, p.10).

O objeto de interesse deste estudo são as ontologias de domínio descritas em linguagens com um grau de formalismo mínimo que permita o processamento dos conceitos, relações, regras e restrições lógicas de uma determinada área do conhecimento por aplicações computacionais.

Considerando as definições apresentadas até o momento espera-se que o conceito de ontologia esteja claro e bem definido na mente do leitor, no entanto, acredita-se ser necessário mostrar um exemplo de uma ontologia real, um objeto concreto.

A Figura 3 ilustra um trecho do conteúdo da *Wine Ontology*[2], uma ontologia desenvolvida por Noy e McGuinness (2001) como um exemplo dentro de um guia para criação de ontologias escrito pelas autoras no ano de 2001 na universidade de Stanford, seu desenvolvimento se deu utilizando o editor de ontologias Protégé e a linguagem OWL.

Figura 3 – Trecho do código fonte da *Wine Ontology*



Fonte: Adaptado de Noy e McGuinness (2001).

Como se pode observar na Figura 3, uma ontologia é essencialmente um arquivo de texto que contém classes compostas por atribuitos, propriedades restritivas, hierárquicas entre outras. É possível visualizar seu conteúdo em um editor de texto, no entanto existem ferramentas específicas para criação, edição e visualização de ontologias.

[2] A *Wine Ontology* é uma ontologia de exemplo criada por Noy e McGuinness (2001) como parte de um guia sobre a criação de ontologias. Utilizando a linguagem OWL DL a *Wine Ontology* pode ser obtida no guia oficial da linguagem OWL. Disponível em: http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine.rdf. Acesso em: 24 jun. 2021.

Dentre estas ferramentas encontra-se o Protégé[3], um editor de ontologias criado pelo departamento de informática médica da Universidade de Stanford. Muito popular no campo de estudos da Web Semântica ele é um *software* gratuito e de código aberto (*open source*), que suporta vários formatos como RDF/RDFS e OWL. (SLIMANI, 2015).

As siglas RDF (*Resource Description Framework*) e RDFS (*RDF Schema*) referem-se respectivamente a um modelo de dados para representação de informações sobre recursos da Web (KLYNE; CARROLL, 2004) e a um vocabulário de modelagem de dados em RDF, utilizado para descrição dos relacionamentos entre os recursos (BRICKLEY; GUHA, 2014). A OWL (*Web Ontology Language*) é uma linguagem projetada para processar o conteúdo da informação ao invés de apenas apresenta-la (MCGUINNESS; HARMELEN, 2004).

Essencialmente uma linguagem para construção de ontologia é um meio para especificar o conhecimento de forma abstrata, isto é, em nível conceitual o que é necessariamente verdadeiro no domínio de interesse. Ela deve ser capaz de expressar restrições que declaram o que deve necessariamente conter qualquer instanciação concreta possível do domínio (ANTONIOU; FRANCONI; HARMELEN, 2005).

Carlan (2006) explica que as linguagens que se prestam a construção de ontologias são na verdade códigos que tem por objetivo principal serem lidos e compreendidos por agentes de *software*, programas capazes de fazerem uso das informações e conhecimentos contidos na ontologia.

Desta forma a escolha de uma linguagem deve levar em consideração aspectos como a capacidade de expressar axiomas, declarações, o uso de operadores e restrições lógicas necessários ao contexto do problema ao qual a ontologia será parte da solução. O domínio da equipe e familiaridade também devem ser considerados, por isso, linguagens já consolidadas para este fim podem ser uma escolha acertada.

Uma variedade de linguagens para criação de ontologias foi desenvolvida nos últimos anos, impulsionadas pela ascensão das pesquisas sobre o campo da Web Semântica e da inteligência artificial. No entanto o W3C (*World Wide Web Consortium*) recomenda a utilização da OWL, que faz parte da pilha de tecnologias recomendas para a Web Semântica.

[3] Disponível em: http://protege.stanford.edu. Acesso em: 21 nov. 2019.

3.3 Web Ontology Language (OWL)

A *Web Ontology Language* (OWL) é linguagem projetada para representar conhecimento rico e complexo sobre coisas, grupos de coisas e relações entre elas (OWL WORKING GROUP, 2012). Sua primeira versão foi desenvolvida pelo *OWL Working Group, um* grupo de pesquisadores ligados ao W3C cuja missão foi produzir recomendações e boas práticas no uso da linguagem (WORLD WIDE WEB CONSORTIUM, 2013).

Santarém Segundo (2010) afirma que OWL é a principal linguagem pra construção de ontologias. Seu principal objetivo é atender às necessidades da Web Semântica e ser efetivamente utilizada por aplicações que necessitem processar o conteúdo das informações e não somente apresentar uma visualização delas.

Em uma pesquisa comparativa com quatro linguagens utilizadas para representar ontologias: KIF, OWL, RDF + RDFS e DAML+OIL, Kalibatiene e Vasilecas (2011) concluem que a OWL é a mais popular e difundida linguagem para Web Semântica e que ainda conta com a maior comunidade.

Para Hogan (2020) a linguagem OWL expande o potencial de expressividade das linguagens RDF e RDFS, com uma semântica mais rica e possibilidades de integração de dados de várias fontes. A segunda versão da linguagem (OWL 2) é recomendada pelo W3C para representar conhecimento sobre coisas, grupos de coisas e seus relacionamentos (OWL WORKING GROUP, 2012).

De acordo com a documentação oficial a linguagem é baseada em lógica computacional, onde o conhecimento é expresso de forma que os algoritmos possam explorá-lo e verificá-lo sistematicamente, possibilitando que informações implícitas sejam tornadas explícitas por meio de inferências. Junto da OWL fazem parte da pilha de tecnologias para Web Semântica ainda o RDF, RDFS e SPARQL (OWL WORKING GROUP, 2012).

A linguagem OWL fornece três sub-linguagens, projetadas para uso distintos:

- *OWL Lite*: indicada quando há necessidade de uma hierarquia de classificação e restrições simplificadas. Devido ao menor conjunto de funcionalidades, é mais fácil de ser utilizada. Uma de suas indicações de uso é para migração rápida de tesauros e outras taxonomias;
- *OWL DL*: *Description* Logic, nome dado devido à sua correspondência com a lógica descritiva. Esta versão tem por característica fundamental a máxima

expressividade da linguagem sem perder a “completude computacional” (todas as conclusões são garantidas como computáveis) e a capacidade de decisão (todas as computações são finalizadas em um tempo finito) dos sistemas de raciocínio. A OWL DL inclui todos os construtores da linguagem, porém eles podem ser usados apenas seguindo certas restrições como a separação entre tipos (uma classe não pode ser ao mesmo tempo um indivíduo ou tipo, e uma propriedade não pode ser ao mesmo tempo um indivíduo ou uma classe);

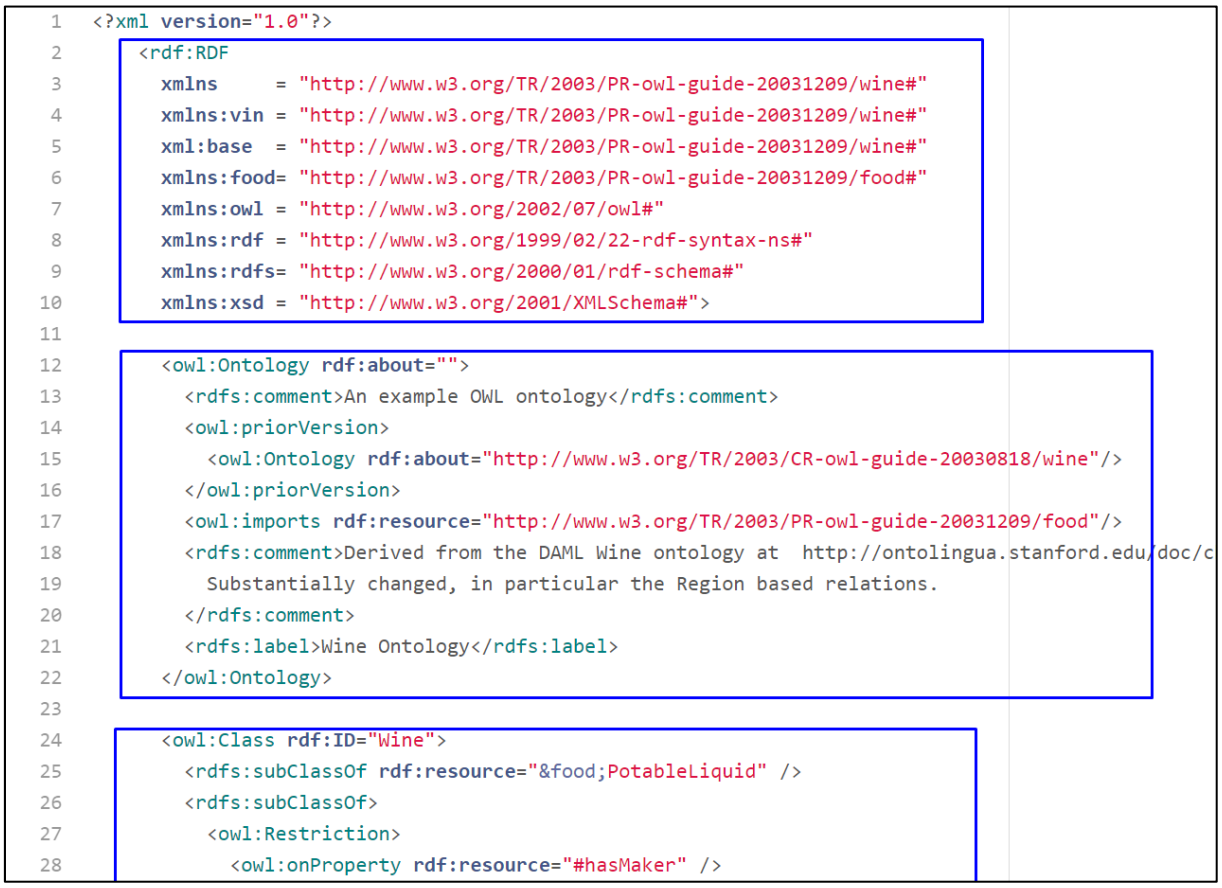
- *OWL Full*: é a versão completa, com o máximo de expressividade e liberdade sintática do RDF, sem garantias computacionais. Esta versão não possui as restrições da OWL DL por isso é indicada para situações onde o aspecto mais importante é a expressividade.

Cada uma destas sub-linguagens é uma extensão de sua antecessora mais simples e a escolha de uma ou outra depende diretamente da análise prévia do domínio e do tipo de ontologia que se pretende criar (SANTARÉM SEGUNDO, 2010).

Os responsáveis pelo desenvolvimento de uma ontologia devem escolher a versão de acordo com as necessidades, levando em consideração que é preciso balancear o nível de expressividade com a capacidade e eficiência de informações. Isto porque, quanto maior o nível de expressividade empregado na ontologia maior será a complexidade dos algoritmos e o processamento computacional necessário nos processos de inferência e busca de informações (HORTÊNCIO-FILHO; LÓSCIO, 2009).

A Figura 4 apresenta um trecho da *Wine Ontology*, que é utilizada como exemplo no guia linguagem da OWL na documentação disponibilizada pelo W3C. O arquivo com extensão “rdf” é formado por alguns “blocos”, destacados com os quadros em azul, onde cada um tem um papel bem definido na estrutura da ontologia.

Figura 4 – Exemplo de um arquivo de ontologia em linguagem OWL



```
1  <?xml version="1.0"?>
2    <rdf:RDF
3      xmlns     = "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
4      xmlns:vin = "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
5      xml:base  = "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
6      xmlns:food= "http://www.w3.org/TR/2003/PR-owl-guide-20031209/food#"
7      xmlns:owl = "http://www.w3.org/2002/07/owl#"
8      xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
9      xmlns:rdfs= "http://www.w3.org/2000/01/rdf-schema#"
10     xmlns:xsd = "http://www.w3.org/2001/XMLSchema#">
11
12     <owl:Ontology rdf:about="">
13       <rdfs:comment>An example OWL ontology</rdfs:comment>
14       <owl:priorVersion>
15         <owl:Ontology rdf:about="http://www.w3.org/TR/2003/CR-owl-guide-20030818/wine"/>
16       </owl:priorVersion>
17       <owl:imports rdf:resource="http://www.w3.org/TR/2003/PR-owl-guide-20031209/food"/>
18       <rdfs:comment>Derived from the DAML Wine ontology at  http://ontolingua.stanford.edu/doc/c
19         Substantially changed, in particular the Region based relations.
20       </rdfs:comment>
21       <rdfs:label>Wine Ontology</rdfs:label>
22     </owl:Ontology>
23
24     <owl:Class rdf:ID="Wine">
25       <rdfs:subClassOf rdf:resource="&food;PotableLiquid" />
26       <rdfs:subClassOf>
27         <owl:Restriction>
28           <owl:onProperty rdf:resource="#hasMaker" />
```

Fonte: Adaptado de Smith, Welty e McGuinness (2004)

Na sequência faz-se uma breve descrição de cada um destes blocos utilizando a linguagem OWL para exemplificar e explicar cada uma das partes mais importantes de uma ontologia.

Namespaces

Geralmente o início de uma ontologia OWL é formado pela declaração dos *namespaces*, responsáveis pela definição dos vocabulários padronizados que serão utilizados dentro da ontologia. Sua função é permitir que os termos possam ser descritos de forma a possibilitar uma interpretação inequívoca e sem ambiguidades.

Na prática os *namespaces* são declarados como atributos de uma *tag* RDF, que por sua vez encontra-se dentro de uma *tag* XML. Neste contexto a linguagem XML permite aos usuários criar documentos com uma estrutura arbitrária (PRADO, 2005). Na Figura 4 as linhas 2 a 10 mostram o uso dos atributos “xmlns” para definição dos *namespaces* utilizados na *Wine Ontology*.

É possível ver na linha 3 que o atributo xmlns não possui a indicação de um prefixo (como os prefixos “vin”, “base” e “food” nas linhas subsequentes), isto faz com que todos os nomes sem prefixo terão como domínio a própria ontologia. Já as linhas 4 e 5 definem respectivamente que os prefixos “vin” e “base” fazem referência a própria ontologia.

A linha 6 define que o prefixo “*food*”, que se refere a uma ontologia já existente, indicada pela URI após o símbolo de igual. Esta é uma forma de reuso de informações pois permite reutilizar, expandir ou mesmo especializar ontologias existentes para atender a demandas específicas.

Por fim as linhas 7 a 10 são os prefixos “owl”, “rdf', “rdfs” e “xsd”, obrigatórios para construção de ontologias usando a linguagem OWL, visto que eles indicam os respectivos léxicos de sintaxe referenciados pelas URLs de cada um.

A utilização destes prefixos tem como principal objetivo facilitar o desenvolvimento da ontologia, abreviando a declaração completa das URLs e permitindo o uso apenas dos prefixos na definição de classes, indivíduos ou propriedades, o que resulta em maior clareza no código e consequentemente mais facilidade na leitura e construção das ontologias.

Cabeçalho

O cabeçalho é a parte onde são especificadas as informações sobre a ontologia, como o nome, comentários e versão, que por sua vez possibilita a referência a versões anteriores desta ontologia. Além disso é dentro do cabeçalho que é feita indicação de uso de código pré-existente, ou seja, a importação outras ontologias, como pode ser visto na linha 17.

Todas estas informações devem são agrupadas dentro da *tag “owl:Ontology”*. Porém é importante esclarecer uma diferença entre a importação de uma ontologia usando a *tag “owl:imports”* e a indicação de uma ontologia em um *namespace*. No primeiro caso o conteúdo completo da ontologia importada é inserido no documento, já o uso do *namespace* permite que apenas trechos da outra ontologia sejam utilizados.

Classes

O terceiro bloco destacado na Figura 4 é o início do que é chamado de corpo da ontologia, que é composto não apenas por classes, mas também por propriedades, indivíduos, relacionamentos, características e restrições (SMITH; WELTY; MCGUINNESS, 2004).

Lima e Carvalho (2005) afirmam que as classes proporcionam um mecanismo de abstração para agrupar recursos com características similares, ou seja, uma classe define um grupo de indivíduos que compartilham algumas propriedades.

A função de uma classe é definir um conceito em um determinado domínio, tal como pessoas, objetos, automóveis ou qualquer outra entidade concreta ou abstrata que se deseja representar. As classes e seus relacionamentos hierárquicos são os elementos básicos para a criação da taxonomia que serve de alicerce para a construção de qualquer ontologia.

Muitos dos usos de uma ontologia dependerão da capacidade de raciocinar sobre os indivíduos. Para isso é necessário descrever as classes as quais os indivíduos pertencem e as propriedades específicas que eles herdam em virtude desta associação (SMITH; WELTY; MCGUINNESS, 2004).

Na linguagem OWL a classe principal e mais genérica (superclasse) da qual derivam todas as demais é chamada de *owl:Thing*, cada classe definida pelo usuário é implicitamente uma subclasse dela.

A Figura 5 mostra um trecho de código da *Wine Ontology* onde é criada a classe "*Wine*", observe que na linha 1 é criada a classe utilizando a *tag owl:Class* seguida do atributo identificador *rdf:ID* com uma identificação única (um nome) da classe entre aspas.

Figura 5 – Exemplo da definição de um Classe OWL

```
1  <owl:Class rdf:ID="Wine">
2      <rdfs:label xml:lang="en">wine</rdfs:label>
3      <rdfs:label xml:lang="fr">vin</rdfs:label>
4      <rdfs:comment>an alcoholic drink made from fermented grape juice.</rdfs:comment>
5      <rdfs:subClassOf rdf:resource="&food;PotableLiquid" />
6      <rdfs:subClassOf>
7          <owl:Restriction>
8              <owl:onProperty rdf:resource="#hasMaker" />
9              <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
10         </owl:Restriction>
11     </rdfs:subClassOf>
12     <rdfs:subClassOf>
13         <owl:Restriction>
14             <owl:onProperty rdf:resource="#hasMaker" />
15             <owl:allValuesFrom rdf:resource="#Winery" />
16         </owl:Restriction>
17     </rdfs:subClassOf>
18
```

Fonte: Adaptado de Smith, Welty e McGuinness (2004)

Para referenciar a esta classe utiliza-se o caractere “#” seguido do nome da classe. No exemplo, para se referir à classe “*Wine”* utiliza-se a declaração “*#Wine*”. Mais especificamente, para se referir a um recurso é necessário o uso da *tag rdf:Resource*. Sendo assim, o código completo seria “*rdf:Resource= "#Wine "”*. Observe por exemplo as linhas 8 e 15 onde são referenciados recursos (ou classes) pertencentes à ontologia.

Outra característica importante de uma ontologia é sua estrutura taxonômica, na Figura 5 nas linhas 5, 6 e 12 pode-se observa o uso da *tag rdfs:subClassOf*, que permite a definição de uma hierarquia entre as classes (conceitos) referenciadas.

Por meio deste recurso é possível definir que uma classe pertence a outra mais genérica, possibilitando que algoritmos compreendam as relações taxonômicas entre conceitos e sejam capazes de fazer generalizações e especificações de forma automática.

Outro exemplo da definição destes relacionamentos de especificidade e generalização pode ser visto na Figura 6, onde são definidas as classes fictícias “Computador”, “*Desktop*”, “*Notebook*”, “*AllInOne*” e “*Netbook*”.

Figura 6 – Exemplo da definição de uma hierarquia de classes em OWL

```
1
2  <owl:Class rdf:ID="Computador"/>
3      <rdfs:label>Computador</rdfs:label>
4  </owl:Class>
5
6  <owl:Class rdf:ID="Desktop">
7      <rdfs:label>Desktop</rdfs:label>
8      <rdfs:subClassOf rdf:resource="#Computador"/>
9  </owl:Class>
10
11 <owl:Class rdf:ID="Notebook">
12     <rdfs:label>Notebook</rdfs:label>
13     <rdfs:subClassOf rdf:resource="#Computador"/>
14 </owl:Class>
15
16 <owl:Class rdf:ID="AllInOne">
17     <rdfs:label>All in One</rdfs:label>
18     <rdfs:subClassOf rdf:resource="#Desktop"/>
19 </owl:Class>
20
21 <owl:Class rdf:ID="Netbook">
22     <rdfs:label>Netbook</rdfs:label>
23     <rdfs:subClassOf rdf:resource="#Notebook"/>
24 </owl:Class>
25
```

Fonte: Adaptado de Pickler (2014, p.82)

Neste trecho de código é possível observar que “*Desktop*” e “*Notebook*” são subclasses de “Computador”, isto implica que “*Desktops*” são computadores, e consequentemente herdam todas as suas características, podendo, no entanto, possuir atributos específicos. O mesmo

acontece com os "*Netbooks*" que são tipos muito específicos de "*Notebooks*", mas que conservam características de sua classe mais genérica "Computador".

Explorando estas relações o método proposto pode localizar termos relacionados aos que compõe a expressão de busca, desempenhando assim a expansão da consulta. A quantidade de termos relacionados identificados depende dos níveis de exaustividade e do tipo de expansão desejado (generalização ou especificação).

Outro tipo de relacionamento entre classes que pode ser definido é a equivalência, que faz uso da *tag owl:equivalentClass*. Este recurso permite representar formalmente que duas ou mais classes são equivalentes ou sinônimas, possibilitando consenso entre termos com grafias diferentes e significados idênticos.

A Figura 7 apresenta um exemplo da definição de uma classe "*Laptop*", um termo considerado equivalente a "*Notebook*".

Figura 7 – Exemplo da definição de uma classe equivalente

```
1
2  <owl:Class rdf:ID="Laptop"/>
3      <rdfs:label>Laptop</rdfs:label>
4      <owl:equivalentClass rdf:resource="#Notebook"/>
5  </owl:Class>
6
```

Fonte: Adaptado de Pickler (2014, p.82)

O recurso de equivalência entre classes também é utilizado pra expansão dos termos de busca no método proposto, porém o foco é a identificação de termos mais significativos para as expressões de busca, promovendo assim resultados mais relevantes.

Axiomas

Além dos conceitos que descrevem um determinado domínio as ontologias são compostas por estruturas hierárquicas, que descrevem as relações entre eles e por axiomas, que possibilitam a definição de restrições semânticas nos conceitos, indivíduos ou propriedades (MIZOGUCHI, 2004; ISOTANI; BITTENCOURT, 2015).

Por meio da utilização dos axiomas, pode-se fazer inferências sobre as entidades, deduzindo e gerando novos conhecimentos a partir das ontologias (ISOTANI; BITTENCOURT, 2015). Sendo assim, os axiomas são basicamente regras que expandem o potencial de expressividade das ontologias, possibilitando que os algoritmos façam deduções lógicas sobre um determinado domínio.

Santarém Segundo e Coneglian (2016, p.222), afirmam que “a inferência é citada em grande parte das pesquisas que tratam de Web Semântica como sendo o grande diferencial na construção de ambientes semânticos”. Os autores ainda explicam que o processo de inferência sobre os dados além de gerar novas informações pode detectar possíveis inconsistências.

Existem diversos tipos de axiomas, de expressão de classes, hierárquicos, de equivalência, cardinalidade, intersecção, complemento, axiomas de propriedades de objetos e de dados além de conectivos booleanos e de enumeração de indivíduos (BECHHOFER *et al*., 2004).

Para explorar todo o potencial das regras e axiomas são utilizados os motores de inferência, que são mecanismos capazes de atuarem seguindo lógicas pré-definidas. Dentre os motores de inferência, que no âmbito da Web Semântica podem ser chamados de *Semantic Reasoners*, destacam-se ELK *Reasoner*, *FaCT++ Reasoner*, *Hermit Reasoner*, O*ntop Reasoner* e *Apache Jena* (SANTARÉM SEGUNDO; CONEGLIAN, 2016, p.225).

Considere o exemplo apresentado por Silva (2012), em uma ontologia existem as regras *hasParent(A, B), hasBrother(B, C)* e *hasUncle(A, C).* Esta ontologia possui ainda um indivíduo chamado João, que tem como pai Mário, que por sua vez é irmão de Alberto. Estes dois fatos estão explicitamente declarados e descritos em seus respectivos indivíduos. No entanto, um motor de inferências que analise esta ontologia é capaz de deduzir que Alberto é tio de João, por meio de inferências a partir do axioma semântico *hasUncle.*

Santarém Segundo e Coneglian (2016, p.228) explicam algumas das propriedades da linguagem OWL:

- “*owl:equivalentProperty*”: indica que duas propriedades são equivalentes quando possuem o mesmo significado;
- “*owl:inverseOf*”: indica uma relação de inversão, se um determinado elemento A se relaciona com o elemento B através de uma propriedade P, e se essa propriedade P é inversa de P2, isso implica que o elemento B se relaciona com A através de P2;
- “*owl:TransitiveProperty*”: indica que se um elemento A tem uma relação X com o elemento B e B tem uma relação X com o elemento C, logo A tem uma relação X com o elemento C;

- "*owl:SymmetricProperty*": indica uma relação de simetria entre dois elementos, indicando que se um elemento A está relacionado a B, isso implica que B também está relacionado a A.

A utilização de axiomas pode incrementar significativamente o potencial de representação semântico de um corpo de conhecimentos, no entanto, o método proposto nesta tese explora apenas as regras hierarquia e equivalência entre classes, deixando a possibilidade de utilização de axiomas sobre objetos e propriedades como uma possibilidade de pesquisas futuras.

Indivíduos

As classes especificam propriedades e características comuns a um conceito, ao passo que os indivíduos são instâncias de uma classe, objetos únicos no mundo real. Em uma ontologia um indivíduo deve obrigatoriamente membro de uma classe.

A documentação da linguagem OWL afirma que indivíduos são instâncias de classes e as propriedades podem ser usadas para relacionar um indivíduo a outro (SMITH; WELTY; MCGUINNESS, 2004). Por exemplo, um indivíduo chamado "Débora" pode ser descrito como uma instância da classe "Pessoa" e por meio da propriedade "trabalhaEm" pode ser ligada ao indivíduo "Universidade de Stanford".

A Figura 8 apresenta um trecho de código onde é definido um indivíduo que é uma instância da classe "*Notebook*". Observe que além do rótulo a outra característica definida no exemplo é o número serial, pois este é um número único de cada indivíduo no mundo real.

Figura 8 – Exemplo da definição de um indivíduo

```
2
3  <Notebook rdf:ID="MeuComputador">
4      <rdfs:label>Meu Computador</rdfs:label>
5      <SerialNumber>71501658159</SerialNumber>
6  </Notebook>
7
```

Fonte: Adaptado de Pickler (2014, p.87)

Propriedades

Utilizando classes e indivíduos é possível definir estruturas taxonômicas bastante complexas. No entanto, para definir fatos gerais sobre classes e específicos sobre indivíduos são utilizadas as propriedades, que são essencialmente relações binárias.

A linguagem OWL dispõem de duas categorias principais de propriedades:

- Propriedades de objetos: estabelecem relação entre classes ou indivíduos;
- Propriedades de dados: indicam a relação entre indivíduos e valores de dados, estes valores devem ser expressos em RDF utilizando os tipos definidos no XML Schema

Na linguagem OWL todas as propriedades são subclasses da classe *rdf:Property*, sendo que para propriedades de objetos usa-se a *tag owl:ObjectProperty* e para as propriedades de dados a *tag owl:DatatypeProperty*. Na Figura 9 pode-se ver a definição de um exemplo de propriedade de objeto denominada "Marca" que está associada à classe computador.

Figura 9 – Exemplo da definição de propriedade de objetos

```
1
2  <owl:ObjectProperty rdf:ID="Marca">
3      <rdfs:label>Marca</rdfs:label>
4      <rdfs:domain rdf:resource="#Computador"/>
5      <rdfs:range rdf:resource="#Fabricante"/>
6  </owl:ObjectProperty>
7
```

Fonte: Adaptado de Pickler (2014, p.84)

Observe que na linha 5 é definida uma restrição de que a propriedade marca deve ser obrigatoriamente preenchida com um dos valores da classe "Fabricante". Enquanto nas propriedades de objetos são definidas as relações entre classes. Nas propriedades de dados são definidos valores como números inteiros, positivos, datas, URLs entre outros. Para definir o tipo de dado que uma propriedade deve ter utiliza-se o léxico do *XML Schema Datatypes*[4].

A Figura 10 mostra um exemplo com duas propriedades de dados, "MemoriaRam" e "Nome", que estão associadas respectivamente às classes "Computador" e "Fabricante". Enquanto a primeira somente aceita valores inteiros positivos a segunda pode ser preenchida com uma cadeia de caracteres.

[4] *XML Schema Datatypes* são os tipos de dados primitivos definidos na especificação do *XML Schema*. Disponível em: https://www.w3.org/TR/xmlschema-2. Acesso em: 26 jun. 2021

Figura 10 – Exemplo da definição de propriedade de dados

```
1
2  <owl:DataTypeProperty rdf:ID="MemoriaRam">
3      <rdfs:label>Memória RAM</rdfs:label>
4      <rdfs:domain rdf:resource="#Computador"/>
5      <rdfs:range rdf:resource="&xsd;positiveInteger"/>
6  </owl:DatatypeProperty>
7
8  <owl:DataTypePropertyrdf:ID="Nome">
9      <rdfs:label>Nome</rdfs:label>
10     <rdfs:domain rdf:resource="#Fabricante"/>
11     <rdfs:range rdf:resource="&xsd;string"/>
12 </owl:DatatypeProperty>
13
```

Fonte: Adaptado de Pickler (2014, p.84)

O emprego adequado das duas categorias de propriedades com suas restrições de tipos de dados e relações entre classes faz com que valores aleatórios ou sem coerência não sejam possíveis de serem informados, o que torna ao mesmo tempo as ontologias mais precisas na representação de um domínio e complexas de serem construídas.

Há ainda algumas formas de especializações que podem ser aplicadas às propriedades, com objetivo de torná-las ainda mais precisas por meio de restrições como propriedades transitivas, simétricas, funcionais ou com um valor de cardinalidade mínima ou máxima (HORTÊNCIO-FILHO; LÓSCIO, 2009). No entanto, estas possibilidades não são exploradas pois são irrelevantes ao contexto do trabalho.

Um recurso da linguagem OWL que precisa ser destacado são as propriedades que possibilitam a definição de rótulos e comentários. A função destes elementos é superar uma limitação comum na área da programação: as restrições de uso de caracteres especiais, acentos e espaços.

É comum em diversas linguagens de programação a impossibilidade do uso de caracteres especiais, acentos e espaços em variáveis, identificadores, nomes de procedimentos entre outros. No idioma inglês isso pode não ser problemático, já em idiomas latinos como o português, francês ou espanhol isso pode ser crítico, pois a limitação pode ocasionar confusões ou dificuldade de interpretação exata em palavras ou expressões.

Para resolver este problema a OWL possui as propriedades de rótulos e/ou comentários, que podem ser definidos utilizando as *tags rdfs:label* e *rdfs:comment* respectivamente. De acordo com a documentação oficial o rótulo fornece um nome para a classe legível por humanos, enquanto os comentários possibilitam a adição de anotações adicionais, sendo que as ferramentas que processam as ontologias podem fazer uso do conteúdo de ambas as *tags*.

Na Figura 11 pode-se ver um trecho da ontologia PEDTERM (2013), utilizada no desenvolvimento do protótipo, com destaque à identificação (ID) da classe "*Bacterial_Disease*" e também os rótulos adicionados para possibilitar a busca por palavras no idioma português.

Figura 11 – Trecho de código OWL com uma classe da ontologia Pedterm

```
<owl:Class rdf:ID="Bacterial_Disease">
    <rdfs:subClassOf rdf:resource="#Infectious_Disease"/>
    <rdfs:label xml:lang="en">Bacterial Disease</rdfs:label>
    <rdfs:label xml:lang="es">Enfermedad Bacteriana</rdfs:label>
    <rdfs:label xml:lang="pt">Doença Bacteriana</rdfs:label>
    <NCI_Definition rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        An acute infectious disorder caused by gram positive or gram negative bacteria. Repre
    </NCI_Definition>
    <NICHD_Definition rdf:datatype="http://www.w3.org/2001/XMLSchema#string">_</NICHD_Definit
    <ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">C2890</ID>
    <SYN rdf:datatype="http://www.w3.org/2001/XMLSchema#string"></SYN>
    <Subset_Association1 rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        NICHD Pediatric Terminology
    </Subset_Association1>
    <Subset_Association2 rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        NICHD Newborn Screening Terminology
    </Subset_Association2>
    <NCI_PT rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        Bacterial Infection
    </NCI_PT>
</owl:Class>
```

Fonte: Adaptado de *Pediatric Terminology* (2013)

Em um estudo utilizando um corpus de 306 ontologias OWL disponíveis na Web, Manaf, Bechhofer e Stevens (2010) identificaram um conjunto de sete diferentes estilos usados na definição de identificadores (IDs) de classes, indivíduos e propriedades. A Tabela 1 apresenta os nomes destes estilos e à direita um exemplo de cada um deles.

Tabela 1 – Estilos de Identificadores em Ontologias

Estilo	Exemplo
Camel Case Style	`QueryExpansionOntologyBased`
Underscore Style	`Query_expansion_ontology_based`
Hyphen Style	`Query-expansion-ontology-based`
Hybrid Camel Case + Underscore Style	`QueryExpansion_ontology_based`
Hybrid Camel Case + Hyphen Style	`QueryExpansion-ontology-based`
Hybrid Hyphen + Underscore Style	`Query-expansion_ontology-based`
Single word	`Query`

Fonte: Adaptado de Manaf, Bechhofer e Stevens (2010)

A escolha de algum destes estilos na criação das ontologias fica a critério dos desenvolvedores, pois não há regras rígidas para isso. Existem apenas diretrizes que recomendam, por exemplo, o uso de palavras no singular, sem o uso de caracteres especiais e

no caso de termos compostos algum dos estilos *Camel Case*, *Underscore* ou *Hyphen* (NOY; MCGUINESS, 2001; MONTIEL-PONSODA *et al*., 2011).

No entanto, é importante destacar que a utilização de rótulos e comentários é opcional, facultando aos criadores da ontologia sua utilização. É permitido ainda declarar mais de um rótulo, possibilitando, por exemplo, que sejam especificados dois termos que se referem a um mesmo conceito (classe).

Além disso, é possível também informar nos rótulos, por meio do parâmetro *xml:lang*, qual o idioma daquela informação, possibilitando assim a criação de ontologias multilíngues, que podem ser reutilizadas em diversos idiomas apenas fazendo apenas a tradução dos rótulos.

Um exemplo da definição de vários rótulos para uma única classe pode ser visto nas linhas 3, 4 e 5 da Figura 11, onde são declarados rótulos nos idiomas inglês, espanhol e português para a classe "*Bacterial_Disease*".

Ressalta-se que para o pleno funcionamento do método proposto nesta pesquisa a utilização rótulos é indispensável, pois é por meio deles que será feita a comparação com os termos informados pelo usuário na expressão de busca com objetivo de fazer a contextualização da consulta.

Ponderando sobre a estrutura interna das ontologias e as características inerentes à sua essência pode-se concluir que as ontologias são artefatos complexos e de difícil desenvolvimento, devido a vários aspectos como a necessidade de profissionais especializados, a alta demanda de tempo para sua construção, o volume de trabalho e os custos (aspecto financeiro) que podem ser limitantes ao amplo desenvolvimento de ontologias.

Nesse sentido mesmo com a crescente demanda por técnicas e instrumentos que colaborem com a transferência de informações no âmbito da Web, que podem ser beneficiadas substancialmente com a utilização das ontologias, o volume de ontologias disponibilizadas publicamente ainda é baixo.

Ressalta-se que algumas áreas de conhecimento, como a área da saúde e da genética possuem um número de ontologias disponibilizadas na Web sensivelmente maior, o que na prática pode resultar em pesquisas e melhorias limitadas às respectivas áreas.

Desta forma buscou-se apresentar um breve resumo das principais motivações e benefícios do investimento na criação de ontologias, um esforço que deve ser empreendido em conjunto por pesquisadores de todas as áreas do conhecimento humano.

3.4 Motivações para criação e utilização de ontologias

Uma ontologia pode ser o coração de um sistema baseado em conhecimento, pois é por intermédio dela que o conhecimento de um domínio pode ser representado (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999). Segundo Sowa (2002) a utilização de ontologias possibilita o desenvolvimento de sistemas mais inteligentes, pois o conhecimento é representado formalmente de modo a suportar a extração explícita, e, mais importante, a extração implícita de conhecimento.

Noy e McGuinness (2001) apresentam uma série de motivos para se desenvolver uma ontologia:

a) compartilhar o entendimento comum de uma estrutura de informação entre pessoas e agentes de *software*;

b) possibilitar o reuso do conhecimento de um domínio;

c) criar especificações e definições explícitas de um domínio do conhecimento;

d) analisar o conhecimento de um domínio;

e) separar o conhecimento de um domínio do conhecimento operacional.

Roa-Martínez (2019, p. 149) aborda outras vantagens práticas e complementares aos motivos de Noy e McGuinness (2001), como:

a) a contribuição como um vocabulário baseado em conceitos para representar o conhecimento;

b) a diminuição da ambiguidade durante a interpretação de um recurso informacional;

c) a facilidade do compartilhamento de uma estrutura de informação comum entre pessoas e agentes de *software*;

d) o compartilhamento de um domínio de conhecimento e;

e) a possibilidade da criação de padrões que levem à interoperabilidade.

Observando os motivos apontados pelas autoras pode-se notar um consenso, uma razão essencial para a criação e utilização das ontologias, que pode ser resumida na seguinte afirmação: a utilização de ontologias possibilita a explicitação do conhecimento de um

determinado domínio, o que permite um entendimento comum e o compartilhamento e reuso das informações.

Ainda discorrendo sobre as motivações para a criação e utilização de ontologias Guizzardi (2000, p.51) resume os principais benefícios em três vertentes:

- *comunicação*: as ontologias são ferramentas úteis na comunicação entre pessoas, sob várias formas, acerca de um determinado conhecimento. Elas podem ajudar no raciocínio e compreensão de um domínio do conhecimento e, portanto, atuam como uma referência para a obtenção do consenso numa comunidade profissional sobre um vocabulário técnico;
- *formalização*: devido a sua natureza formal, a especificação de uma ontologia de domínio elimina contradições e inconsistências envolvendo as restrições, resultando assim que uma ferramenta não ambígua. Um outro ponto a ser destacado é que a especificação formalizada pode ser automaticamente verificada e validada, possibilitando assim que mecanismos de inferência sejam capazes de derivar novos conhecimentos de forma automática, a partir da base de conhecimento já presente na ontologia;
- *representação do conhecimento e reuso*: as ontologias formam um vocabulário de consenso que permite representar conhecimento de um domínio em seu nível mais alto de abstração, possuindo assim um alto potencial de reutilização.

Biscalchin e Boccato (2011) em um estudo sobre o uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias, apontaram que a incompatibilidade entre a linguagem da busca e a adotada pelo sistema e as inconsistências sintático-semânticas originam falhas nos sistemas de informação, que provocam grande insatisfação por parte dos usuários. Nesse sentido, as ontologias de domínio podem ser utilizadas como recursos de apoio à interação entre usuários e sistemas de informações.

Desta forma, acredita-se que na vertente comunicacional, a utilização de ontologias de domínio pode mudar significativamente os processos de armazenamento, gestão e recuperação de informação, por meio de um mecanismo de compatibilização terminológica entre a linguagem utilizada para representar as informações disponíveis no acervo e a linguagem empregada pelo usuário na elaboração da expressão e busca.

Outros projetos e soluções em distintas áreas de pesquisa também podem usufruir das vantagens do emprego de ontologias, como por exemplo na gestão do conhecimento, onde Andrade, Ferreira e Pereira (2010), descrevem uma solução aplicada ao processo de desenvolvimento de produto. Neste contexto as ontologias foram utilizadas na classificação e recuperação do conhecimento, promovendo um melhor compartilhamento e difusão dos conhecimentos tácitos e explícitos entre os indivíduos envolvidos no processo de desenvolvimento de produto.

Há ainda outras pesquisas na área da gestão do conhecimento como Barquin, González e Pinto (2006), Lopes (2011), Cardenas (2014), Lourenço, Ramalho e Penteado (2015) e Nascimento e Pinho (2018) que visam organizar e prover algum nível de padronização da informação com vistas a sua melhor recuperação.

No processamento de linguagem natural as ontologias podem ser empregadas na análise do discurso localizando os conceitos centrais em textos em linguagem natural, identificando o contexto correto de termos com a mesma grafia e significados diferentes por meio da análise da sentença ou ainda servindo como ferramenta de apoio a tradução de termos com mais acurácia que um sistema baseado apenas em um dicionário. Tais propostas podem ser observadas respectivamente nos trabalhos de Busch *et al.* (2006) e Aguado *et al.* (1998). Abordagens semelhantes podem ser vistas em Moreno e Pérez (2000), Liu, Hogan e Crowley (2011), Pereira e Rigo (2013) e Harispe *et al.* (2015).

Acredita-se ainda que os projetos que relacionados à recuperação de informação são os mais beneficiados pela utilização de ontologias. Tal conclusão origina-se do volume de trabalhos encontrados nesta temática, como o Textpresso, que é um sistema de mineração de texto para literatura científica que utiliza uma ontologia baseada em classes e subclasses de conceitos biológicos para recuperar documentos científicos (MÜLLER; KENNY; STERNBERG, 2004).

Outra abordagem clássica é a de Castells, Fernandez e Vallet (2006), que apresentam uma proposta de um sistema de recuperação ode documentos baseado em ontologias onde as palavras chave informadas na busca passam por um mecanismo que faz uma expansão por meio de anotações de forma semiautomática, utilizando o modelo vetorial os autores demonstraram melhorias sensíveis nos resultados de busca.

Há ainda abordagens que buscam descobrir o contexto das necessidades informacionais do usuário utilizando ontologias (FERNÁNDEZ *et al.*, 2011), outras baseiam-se na indexação

semântica, como o trabalho de Kara *et al.* (2012), que apresentam um sistema de recuperação de informação que durante a etapa de extração de termos para construção do índice faz uso de ontologias, o que possibilita a inferência automática e a descoberta de novas informações que são agregadas à indexação. A proposta demonstrou ganhos de precisão e revocação por meio da resolução de problemas estruturais e ambiguidades.

Considerando que uma das maiores dificuldades enfrentadas pelo usuário de um sistema não é exatamente a recuperação de documentos, mas a recuperação de informação relevante para seus propósitos de pesquisa, nos documentos recuperados (ARAÚJO, 2012). O uso de ontologias pode promover melhorias significativas, visto que de acordo com Simões *et al.* (2017, p.164) elas contribuem na:

a) desambiguação de termos homógrafos e polissêmicos;

b) capacidade de integração de relações semânticas de forma automatizada;

c) navegação e expansão de consultas através de relações semânticas;

d) recuperação mais precisa e exaustiva da informação.

Dessa forma o método de contextualização e expansão de termos da expressão de busca baseado nas relações semânticas das ontologias proposto neste estudo pode incrementar substancialmente os índices de satisfação das necessidades informacionais dos usuários.

Nos próximos capítulos são abordados detalhadamente os processos de expansão e contextualização de consultas baseados nas relações semânticas das ontologias de domínio.

4 Método de contextualização e expansão de consultas baseado em ontologias de domínio

Este capítulo apresenta o método de contextualização da expressão de busca utilizando a estrutura terminológica das ontologias de domínio e expansão de consulta utilizando os relacionamentos semânticos entre os conceitos das ontologias com objetivo de auxiliar o usuário na elaboração de sua expressão de busca. A proposta almeja aprimorar a interatividade entre o mecanismo de busca e seus utilizadores, proporcionando melhores condições de representar suas necessidades informacionais e consequentemente contribuir para a eficiência do processo de recuperação de informação.

Os sistemas de recuperação de informação Web, também conhecidos como mecanismos de busca, possuem um papel primordial como intermediadores entre a informação registrada e os usuários que a necessitam (SOUZA, 2006, p. 162). Alguns dos mecanismos de busca mais conhecidos são baseados em caixas de texto livre, isto é, utilizam uma dinâmica de funcionamento onde o usuário descreve sua necessidade de informação por meio de um conjunto de termos (expressão de busca) e o sistema utiliza esses termos para efetuar uma comparação com as representações dos documentos de seu acervo. Um documento poderá ser recuperado se a sua representação coincidir total ou parcialmente com a expressão de busca, resultando uma lista de documentos usualmente ordenada pelo grau de coincidência de cada documento em relação à expressão de busca.

A representação ao qual refere-se a proposta é a representação temática, a representação de um conteúdo informacional, dos assuntos. No contexto dos documentos, considerando o arcabouço teórico da área de RI, acredita-se que estas representações podem ser feitas de forma satisfatória. Comumente os documentos são formados por um volume razoável de texto, sendo assim possível inferir seus assuntos por meio de distintos métodos.

O mesmo não se pode dizer da representação da necessidade de informação do usuário. Considerando que as expressões de busca normalmente são formadas por um número reduzido de termos, a interpretação da real necessidade de informação do usuário apenas com base na expressão de busca é uma tarefa complexa e que assume significativa importância.

A maioria dos atuais mecanismos de busca são produtos oferecidos por empresas privadas, que não divulgam abertamente as suas técnicas e tecnologias de busca. Desta forma, as afirmações acerca do funcionamento destes mecanismos são baseadas em testes e conhecimentos empíricos, adquiridos ao longo dos anos de utilização e convivência com tais ferramentas.

Como discutem Beppler (2008) e Baeza-Yates e Ribeiro Neto (2013), o processo de recuperação de informação utilizando mecanismos de busca baseados em caixas de texto pode complexo em razão das dificuldades relacionadas à elaboração das expressões de busca.

Dentre os principais problemas enfrentados pelos usuários, observados por Camacho (2004), Beppler (2008), Baronas (2011) e Nayak, Prasad e Senapati (2015), ressaltam-se as principais dificuldade em: a) expressar suas necessidades informacionais utilizado palavras-chave, b) compreender o domínio pesquisado, c) compreender a terminologia e interpretar as relações entre conceitos do domínio pesquisado e d) lidar com problemas linguísticos como sinonímia, homonímia e polissemia

Este trabalho propõe um método de utilização de ontologias de domínio para facilitar a execução desta atividade. O método baseia-se em duas etapas:

1. *Contextualização da necessidade de informação* do usuário a partir da identificação de coincidências terminológicas entre os termos de sua expressão de busca e os conceitos/termos de um conjunto de ontologias;
2. *Expansão da consulta* por meio da identificação, na ontologia, de termos direta ou indiretamente relacionados aos termos que compõe a expressão de busca inicial, com o propósito de auxiliar o usuário na enunciação e formalização de sua necessidade de informação.

Ambas as etapas são distintas, porém a segunda depende diretamente da primeira. Desta forma, elas devem ser executadas na ordem especificada.

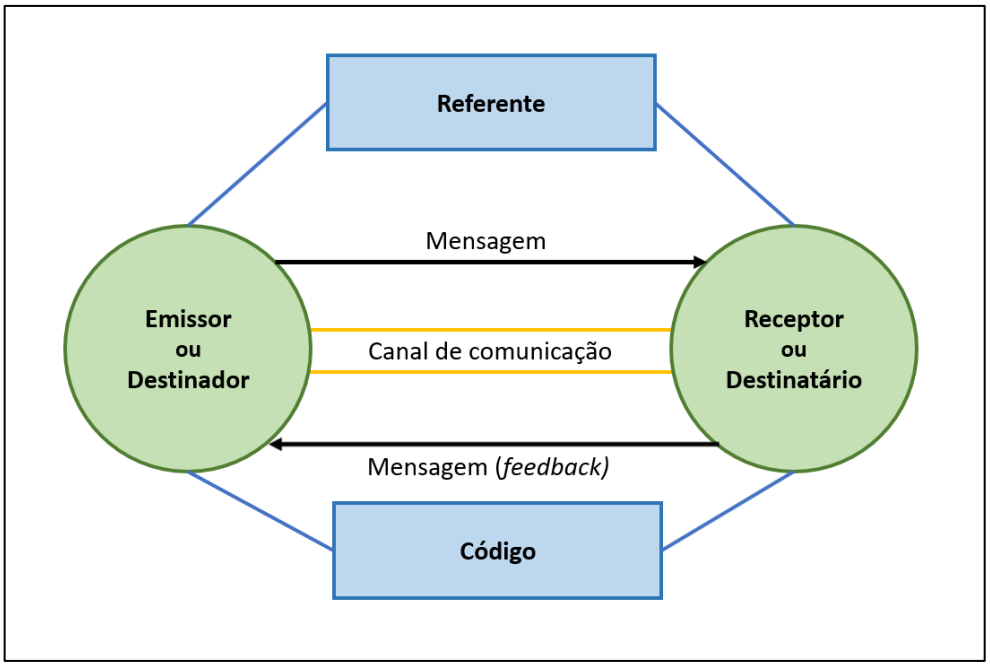
4.1 Contextualização da Necessidade de Informação

Segundo Vieira (1994, p.6), a recuperação de informação é um processo de comunicação em que o emissor e o receptor se relacionam para atender uma necessidade de informação. Ao fazer uma pergunta (consulta) ao sistema, o usuário funciona como emissor e o computador

como receptor. Em contrapartida, o computador, ao apresentar a sua resposta passa a ser o emissor e o usuário o receptor.

A teoria da comunicação explica que todo processo de comunicação tem por objetivo transmitir uma mensagem. De forma geral ele é composto por alguns elementos essenciais: o emissor, o receptor, a mensagem, o canal, o código e referente (VANOYE, 1992). Na Figura 12 é possível observar o relacionamento entre estes elementos.

Figura 12 – Esquema da comunicação



Fonte: Adaptador de Vanoye (1992)

Uma definição sucinta de cada elemento do processo comunicativo é feita com base nas explanações de Vanoye (1992):

- emissor ou destinador: é quem emite a mensagem;
- receptor ou destinatário: é quem recebe a mensagem; pode ser um indivíduo, um grupo ou um computador (ou sistema);
- mensagem: é o objeto da comunicação, constituída pelo conteúdo das informações transmitidas;
- canal de comunicação: é a via de circulação das mensagens;
- código: é o conjunto de signos e regras de combinação destes signos. Pode ser a língua oral ou escrita, gestos, código Morse, etc.;

- referente: é constituído pelo contexto, pela situação e pelos objetos reais aos quais a mensagem remete.

Sobre o elemento *referente*, Amaral (2007) complementa que ele pode se constituir na situação, nas circunstâncias de espaço e tempo em que se encontra o destinador da mensagem ou ainda dizer respeito aos aspectos textuais da mensagem.

Considerando a concepção de Vieira (1994), onde o usuário e o mecanismo de busca se alternam nos papéis de emissor e receptor, é possível transpor os demais elementos para o contexto da recuperação de informação, sendo que:

- **a mensagem**: é a expressão de busca, composta por um ou mais termos que representam a necessidade informacional do usuário;
- **o canal:** é a internet, que permite ao usuário acessar endereço do sítio do mecanismo de busca;
- **o código**: é a língua (o idioma) ao qual pertencem o léxico, os termos escolhidas pelo usuário para construção de sua expressão de busca;
- **o referente:** é todo o contexto do usuário, aspectos físicos, técnicos e cognitivos em relação à sua expectativa frente a mensagem transmitida.

Nesta pesquisa considera-se o contexto como o domínio do conhecimento humano ao qual pertence a mensagem transmitida pelo usuário por meio de sua expressão de busca, pois possibilita ao sistema inferir a significação pretendida pelo usuário.

O dicionário Oxford (OXFORD LANGUAGES, 2021) define significação como

a) a representação mental relacionada a uma forma linguística, um sinal, um conjunto de sinais, um fato, um gesto etc.;

b) aquilo que um signo quer dizer;

c) acepção, sentido, significado.

Desta forma, a contextualização da expressão de busca permite que o sistema seja capaz de inferir ou deduzir o sentido e o significado de cada um dos termos que a compõe, e por consequência se aproxime-se da real necessidade de informação do usuário.

Ontologia como representação do contexto

Como discutido no Capítulo 3, as ontologias são instrumentos que necessariamente incluem um vocabulário de termos e alguma especificação de seus significados. Elas necessitam também de definições e da indicação de como os conceitos estão inter-relacionados, o que impõem uma estrutura no domínio e restringe possíveis interpretações de termos (JASPER; USCHOLD, 1999).

Na definição dos autores é possível observar a utilização dos vocábulos "termos" e "conceitos" como sinônimos. No entanto, este pensamento não é consensual entre pesquisadores de distintas áreas do conhecimento.

Considera-se neste estudo o entendimento de que um conceito, na dimensão do referente, pode ser definido como uma unidade de pensamento. Este raciocínio baseia-se na teoria de Dahlberg (1978), que reconhece o conceito como uma unidade de conhecimento que é construída a partir das relações de afinidade com outros conceitos.

Barros (2016) afirma que o conceito é a unidade de pensamento da realidade observável, enquanto o termo é a palavra utilizada para comunicá-lo. Para Sowa (1984, p. 344) os conceitos são invenções da mente humana usadas para construir um modelo de mundo, enquanto Sultagubiyeva, Raushangul e Gulzhamal (2015, p. 116, tradução nossa, grifo nosso) salientam que:

> O conceito pode ser denominado como uma ideia abstrata [...] associada a uma representação correspondente na linguagem, que denota todos os objetos em uma determinada categoria ou classe de entidades, interações, fenômenos ou relações entre eles [...]. **Os conceitos existem** na mente como entidades abstratas **independentes dos termos usados para expressá-los**.

Já os termos são definidos por Medeiros (1986, p. 2) como

> Signos linguísticos que representam um conceito identificado na estrutura conceitual de um campo específico do conhecimento. Distinguem-se os termos das palavras da linguagem comum pela relação de univocidade (a um conceito corresponde apenas uma denominação) e de monorreferencialidade (a uma denominação corresponde apenas um conceito), que são estabelecidas entre conceito e termo.

Fica claro então que os conceitos são entidades abstratas, intangíveis, que existem na cognição humana, enquanto os termos são as representações físicas e palpáveis dos conceitos.

Para Cabré (1993), as unidades terminológicas têm a capacidade de referenciar desde que estejam contextualizadas, visto que os conceitos de um mesmo âmbito especializado mantêm entre si relações que constituem a estrutura conceitual do domínio estudado. O valor de um termo depende de sua posição relativa no texto, isto porque o texto manifesta o domínio de conhecimento por meio da representação pela escrita independente de seu suporte.

De acordo com Dubuc (1985 *apud* Cervantes, 2020, p.9) a situação em que os termos se encontram transcende a noção comunicacional, assim, uma mesma noção poderá encontrar marcas diferentes dependendo da sua área de utilização, isto é, o termo adquire sua função semântica a partir da sua associação em um domínio específico.

A autora afirma que os contextos são fundamentais, uma vez que ilustram o uso real de um termo, exprimem a ideia completa sobre ele. Sendo assim, "não se pode analisar os termos desvinculados das condições nas quais eles foram produzidos e do tipo de discurso de onde eles foram extraídos" (CERVANTES, 2020, p.9).

Este estudo considera o domínio "como uma parte do saber cujos limites são estabelecidos conforme um ponto de vista particular" conforme define a Norma ISO 1087 (2002). Já o contexto "é uma noção escorregadia, que se mantém na periferia e foge quando se tenta defini-lo" (DOURISH, 2004, p. 29, tradução nossa).

A definição de contexto que aparece associada as pesquisas no campo da RI e dos SRI, conhecida como recuperação de informação baseada em contexto refere-se aos sistemas que levam em consideração as informações sobre

> as intenções do usuário (tarefa a ser realizada, percepção da tarefa, tipo de informação buscada), o próprio usuário (conhecimento, perfil e cultura), seu ambiente (ambiente físico e histórico da tarefa), o campo das necessidades de informação (natureza do corpus e áreas cobertas) e as características do sistema (representação de documentos, método de correspondência de consulta/documento, interface de visualização e estratégias de acesso à informação) (BOURAMOUL, 2016, p. 74, tradução nossa)

O autor ainda afirma que o objetivo da recuperação de informação baseada em contexto é atender melhor às necessidades de informação do usuário, integrando o contexto de pesquisa na cadeia de acesso à informação.

Para Dey (2001), o contexto pode ser definido como qualquer informação que pode ser utilizada para caracterizar a situação de uma entidade, sendo que uma entidade pode ser uma pessoa, lugar ou objeto considerado relevante para a interação entre um usuário e sistema.

Neste sentido, o “contexto” é entendido como o conjunto de dados sobre um usuário em uma busca de informações, abarcando características do ambiente, da tarefa, do dispositivo utilizado, do dia, horário, local, entre outras.

A definição de contexto abordada neste estudo assume um sentido distinto, considera-se como “contexto” o domínio de utilização de um termo, sua modalidade, condições de utilização e elementos semânticos e definitórios que permitem estabelecer a relação entre um termo e o conceito por ele representado (CERVANTES, 2020).

A autora ainda explica que

> não se pode analisar os termos desvinculados das condições nas quais eles foram produzidos e do tipo de discurso de onde eles foram extraídos. Desse modo, os **contextos** são fundamentais uma vez que **ilustram o uso real de um termo, exprimem uma ideia completa sobre ele**. Os contextos devem ser escolhidos em função de suas qualidades e **servem para elucidar a denominação e transmitir com clareza o conceito que o termo representa**. (CERVANTES, 2020, p.10, grifo nosso)

Para Tramontin Junior (2011, p. 75) “o contexto é tido como o significado dos termos, que algumas vezes é obtido através das palavras que se encontram próximas aos termos em questão”, reforçando a noção de contexto como sentido de um ou mais termos.

Considerando que o contexto de um determinado termo é imprescindível à sua plena e inequívoca compreensão e comunicação, entende-se que para possibilitar a um usuário as condições adequadas de exprimir suas necessidades informacionais, em um mecanismo de busca utilizando um conjunto de termos, é preciso possibilitar a definição do contexto de sua busca.

Para que a contextualização de uma busca se torne possível é necessário que as definições e relações semânticas em um determinado domínio do conhecimento estejam formalizadas em algum tipo de suporte, as ontologias são instrumentos adequados a esta função.

Desta forma, o método proposto utiliza as ontologias como instrumentos capazes de representar o contexto de um termo em um determinado domínio do conhecimento. Ou seja, as ontologias são representações do contexto geral de uma busca.

Na prática, quando um ou mais termos da expressão de busca são localizados em uma ontologia de domínio, esta passa a ser uma possível candidata a representar o contexto desta consulta, permitindo ao sistema melhores condições de compreender as necessidades informacionais do usuário.

Entende-se que, se a terminologia de uma determinada área do conhecimento, materializada nas ontologias de domínio, representa o corpo conceitual desta área, então caso um ou mais termos de uma expressão de busca estejam contidos em uma ontologia, a área de conhecimento desta expressão pode ser a mesma da ontologia. Em outras palavras, se os termos de uma determinada expressão de busca aparecem em uma ontologia de domínio, a área de conhecimento desta ontologia pode ser a mesma da expressão de busca, sendo assim esta ontologia representa o contexto da busca.

Considera-se ainda no âmbito terminológico que o contexto de uma busca diz respeito ao que se trata aquela busca, em outros palavras, o domínio da busca. Sendo assim, no decorrer do trabalho “contexto” e “domínio” são vocábulos utilizados como sinônimos, pois remetem à mesma ideia geral.

Cervantes (2020) afirma que os termos adquirem função semântica a partir da sua associação a um domínio específico do conhecimento, um contexto, que por sua vez é fundamental, pois ilustra o uso real de um termo, servindo ainda para elucidar sua denominação e transmitir com clareza o conceito que ele representa.

Desta forma, a definição do contexto é uma etapa necessária a um mecanismo de busca que se preste a auxiliar seus utilizadores a representarem satisfatoriamente suas necessidades informacionais. Outras abordagens, como a expansão da consulta, também dependem do contexto para desempenharem suas abordagens adequadamente.

4.2 Expansão da consulta

Na busca por melhorias no desempenho dos mecanismos de recuperação de informação, mais especificamente no sentido de obter resultados mais relevantes para os usuários, várias abordagens foram propostas ao longo dos anos, cada qual tentando superar dificuldades no processo de RI como um todo.

Uma das vertentes é a melhoria na elaboração das expressões de busca submetidas aos Sistemas de Recuperação de Informação (SRI), elementos mediadores entre os usuários e um determinado acervo documental (FERNEDA, 2013). Nesses mecanismos o usuário interage com o sistema comunicando sua necessidade de informação para que obtenha uma lista de

documentos que possam satisfazer tais necessidades. Em muitos sistemas essa comunicação é feita por meio da especificação termos que representem a necessidade do usuário.

O autor explica que a escolha destes termos deve ser criteriosa, pois eles serão responsáveis por comunicar ao sistema as necessidades informacionais do usuário, possibilitando assim a recuperação de itens relevantes. Isto caracteriza uma situação delicada, pois um usuário que busca informações sobre um determinado domínio do conhecimento pode não ter familiaridade com sua terminologia, ocasionando assim expressões de busca inadequadas e/ou insuficientes.

Outro fator que pode levar a uma expressão de busca ineficiente está relacionado à falta de um conhecimento sobre a coleção de documentos, ocasionando dificuldades em formular consultas capazes de recuperar as informações desejadas. Sendo assim, é comum que os usuários "precisem reformular suas consultas muitas vezes para obter resultados que lhes interessam" (BAEZA-YATES E RIBEIRO-NETO, 2013, p. 157).

Há de se considerar ainda a discrepância entre os termos conhecidos pelo usuário e os empregados na indexação dos documentos, provenientes da linguagem técnica dos autores ou ainda de vocabulários controlados utilizados nos processos de indexação.

Todos estes fatores podem resultar em uma falta de coincidência terminológica, que por sua vez pode fazer com que informações relevantes não sejam facilmente encontradas, tornando o processo de recuperação de informação desnecessariamente difícil ou em determinadas situações até impossível caso não haja ferramentas para auxiliar o usuário.

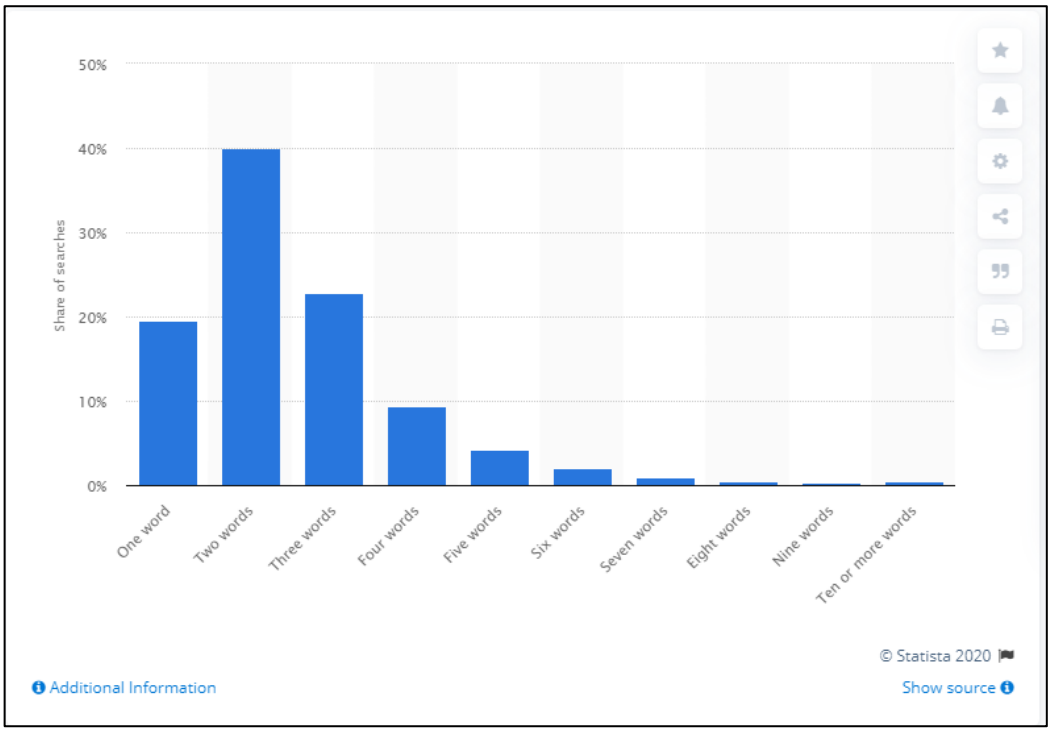
Spink *et al.* (2001) realizaram um estudo com mais de um milhão de consultas no mecanismo de busca "Excite". Foi observado que o número médio de termos utilizados varia entre 2 e 3 e mais da metade dos usuários reformulam suas buscas ao menos uma vez, o que reforça a percepção de que os usuários tem dificuldades em elaborar expressões de busca.

Em um levantamento mais atual sobre a quantidade de termos utilizados em uma consulta foram encontrados alguns indicadores no portal de estatísticas *statista.com*, pertencente à empresa Statista[5], uma fornecedora de dados de mercado para diversas empresas.

[5] Disponível em: https://www.statista.com/. Acesso 02 jun. 2020.

A Figura 13 apresenta um gráfico com o número médio de termos de pesquisa para consultas em mecanismos de busca nos Estados Unidos para o mês de janeiro de 2020. Ele foi extraído de um dos indicadores do portal *statista.com* elaborado por Johnson (2021).

Figura 13 – Número médio de termos de pesquisa em consultas nos Estados Unidos - Janeiro de 2020



Fonte: Adaptado de statista.com (2020)

É possível notar que mais de 80% das buscas tem três ou menos palavras, sendo a maior parcela, 40% de apenas duas. Além disso, menos de 5% de todas as buscas são compostas por cinco ou mais termos, um indicativo de que os usuários não elaboram consultas complexas ou detalhadas, confirmando a taxa de 50% de reformulações observada por Spink *et al.* (2001).

Outra pesquisa que utilizou logs de utilização de um mecanismo de busca Web foi a de Rieh e Xie (2001), na qual os autores concluem que os usuários nem sempre sabem como expressar o que estão procurando em uma consulta, pois, com base nos resultados obtidos para que as buscas atendessem aos objetivos iniciais foram necessárias sucessivas reformulações na consulta original.

Joachims *et al.* (2007) descobriram que os usuários faziam em média 2,2 reformulações para responder a uma única questão, o que reforça a percepção de que as consultas iniciais elaboradas pelos usuários muitas vezes não são suficientes para localizar informações satisfatórias, forçando-os a tomar alguma medida.

Considerando as pesquisas de Rieh e Xie (2001), Jones *et al*. (2006) e Xiang *et al*. (2010) pode-se resumir as motivações que levam os usuários a reformularem suas consultas em cinco categorias:

- substituição por sinônimos: quando alguns dos termos originais são alterados por palavras que tenham o mesmo significado;
- correção de erros: ocorre quando o usuário nota que cometeu algum erro ortográfico ou de digitação, no caso de palavras homófonas a reformulação também é considerada uma correção de erros;
- especificação: quando se deseja um conjunto de resultados mais específico, é feita a substituição de alguns termos ou a adição de novos com significado mais específico;
- generalização: quando se deseja um conjunto de resultados mais genérico, é feita a substituição de alguns termos com significado mais geral ou ainda a exclusão de termos com significado específico;
- substituição por termos relacionados: diferente da substituição por sinônimos, aqui há uma mudança de significado. No entanto, os novos termos adicionados ou alterados possuem alguma relação com os termos originais.

Estas dificuldades indicam que a elaboração de consultas é frágil e complexa, sendo necessário um suporte por parte do sistema para que os usuários tenham condições de representar melhor suas necessidades de informação e consequentemente obtenham resultados mais relevantes.

Motivada pela importância e pelas dificuldades no processo de especificação de buscas surgiu dentro da área de Recuperação da Informação um nicho de pesquisa em expansão de consulta (*query expansion*), caracterizado pelo estudo e métodos e processos que visam melhorar a eficiência da recuperação de informação por meio do auxílio a construção de consultas que melhor reflitam as necessidades de informação dos usuários (FERNEDA, 2013, p.94). O autor ainda afirma que a expansão de consultas está relacionada com o conceito mais genérico de reformulação de consultas, que pode envolver também a exclusão de termos do conjunto inicial.

Bhogal, Macfarlane e Smith (2007) afirmam que o principal objetivo da expansão da consulta é adicionar novos termos significativos à consulta original. Porém, é importante que

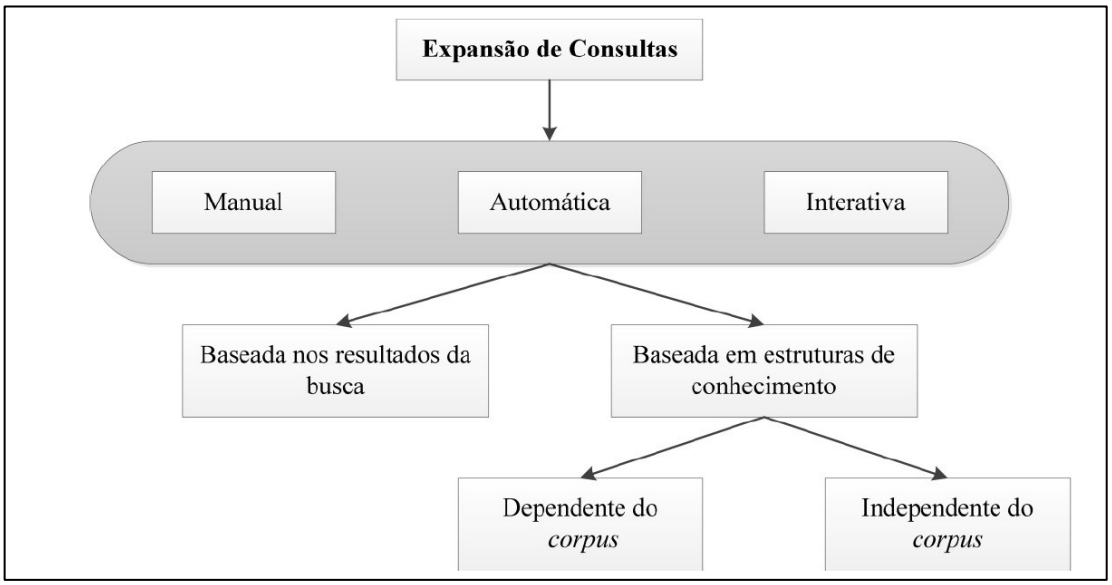
os termos a serem adicionados reflitam as reais intenções do usuário, resultando assim em uma expressão de busca que represente com maior precisão as necessidades informacionais do usuário.

Para Efthimiadis (1996), o processo de busca é dividido em duas etapas: a formulação da consulta inicial, onde é definida a estratégia de busca, e, após uma breve análise dos resultados pode ser necessária a segunda etapa: a expansão da consulta, que pode ser feita manualmente ou com intervenção do sistema, resultando assim um conjunto de documentos mais relevantes.

De acordo com Efthimiadis (1996) e Bhogal, Macfarlane e Smith (2007), há essencialmente três modos de expansão de consulta: *manual*, *automática* ou assistida pelo usuário ou *interativa*. A forma *manual* depende exclusivamente do usuário, sendo o único responsável por reescrever a consulta, decidindo quais termos serão incluídos ou substituídos e guiado apenas por seus conhecimentos. Na expansão *automática* o sistema é responsável pela geração e escolha dos termos que irão compor a nova consulta. Há ainda uma forma intermediária em que o sistema gera um conjunto termos de expansão e os apresenta ao usuário, que por sua vez escolhe qual destes irá incorporar a expressão de busca.

A Figura 14 apresenta um diagrama com as três formas supracitadas de se fazer a expansão das consultas e os métodos de geração dos novos termos utilizados no processo.

Figura 14 – Tipos e abordagens para expansão da consulta



Fonte: Adaptado de Efthimiadis (1996)

Em relação às formas de obtenção dos termos utilizados na expansão das consultas é possível agrupar as abordagens em dois grupos: baseadas nos resultados da busca ou baseadas em estruturas de conhecimento. Nos métodos baseados nos resultados da busca os termos para expansão são obtidos a partir dos documentos recuperados após a consulta inicial. Neste caso Ferneda (2013) afirma que a eficácia da expansão depende fortemente da qualidade da consulta inicial, limitação que não existe nos modelos baseados em estruturas de conhecimento.

As abordagens baseadas em estruturas de conhecimento também dividem-se em duas vertentes: dependentes ou independentes do corpus. Nos primeiros todo o acervo é analisado para extração dos termos que serão posteriormente utilizados para a expansão das consultas.

Já os métodos independentes fazem uso de estruturas de conhecimento que não possuem relação com o acervo, como por exemplo os glossários, dicionários, tesauros e ontologias. Bhogal, Macfarlane e Smith (2007) observa que neste tipo de abordagem é imprescindível que os modelos de conhecimento sejam processáveis por máquinas, como as ontologias, para que os conceitos que eles definem possam ser utilizados por soluções automáticas.

As abordagens de expansão de consulta baseadas nos resultados da busca possuem estreita relação com o conceito de *Relevance Feedback*, pois partem da premissa de que embora seja complexo formular uma primeira expressão de busca eficiente, é fácil julgar a relevância dos documentos recuperados.

De acordo com Ferneda (2013) embora os métodos de *Relevance Feedback* provem ser eficazes na melhoria dos resultados da recuperação eles possuem alguns inconvenientes, como a impossibilidade de uso em caso de poucos ou nenhum resultado de uma busca inicial e a dependência da voluntariedade dos usuários em fornecer informações sobre a relevância os documentos recuperados.

Considerando o comportamento da sociedade, no qual as pessoas buscam formas de satisfazer suas necessidades de informação de forma cada vez mais rápida e imediata, soluções que demandam interações como a leitura e análise de um conjunto mínimo de resultados para possibilitar uma nova busca mais relevante tornam-se práticas com menor potencial de aceitação.

Desta forma as técnicas que demandam menor envolvimento do usuário e que possam ser utilizadas já na busca inicial podem ser mais promissoras, como os métodos de expansão de consultas independentes do *corpus*.

Abordagens baseadas em estruturas de conhecimento independentes do *corpus*

Para Ferneda (2013) as estruturas de conhecimento constituem fontes especialmente promissoras para a geração dos termos de expansão, pois, independentemente do tamanho do acervo, possibilitam a expansão da consulta inicial a qualquer momento.

Explorando as relações semânticas entre palavras, algumas pesquisas fizeram uso do WordNet, um banco lexical que se assemelha a um dicionário de sinônimos. Composto por substantivos, verbos, adjetivos e advérbios agrupados em conjunto de sinônimos cognitivos (chamados de *synsets*) cada um expressando um conceito distinto. Além disso, estes *synsets* são interligados por meio de relações conceituais-semânticas e lexicais, possibilitando assim uma rede de conceitos semanticamente desambiguadas (WORDNET, 2020).

Gong, Cheang e Hou (2005) propuseram um método de expansão de consultas na Web utilizando o WordNet para expansão de consultas nas dimensões de hiperonímia, hiponímia e sinonímia. A co-ocorrência entre termos também foi utilizada para reduzir ruídos e complementar a abordagem. Os autores observaram resultados positivos no desempenho de consultas na Web.

No trabalho de Zhang, Deng e Li (2009) a expansão de termos é feita em duas etapas, inicialmente a sentença da consulta original passa por uma desambiguação de sentido utilizando o módulo de medida de relacionamento semântico da WordNet, na sequência os conceitos recuperados são expandidos via banco de dados lexical, produzindo consultas expandidas que demonstraram 7% de melhoria na precisão em relação a métodos convencionais.

Existem ainda abordagens que fazem uso de tesauros, como a técnica de Mandala, Tokunaga e Tanaka (1999) que combina diferentes tipos de tesauros para efetuar a expansão de consultas e obter melhores resultados na recuperação de informação.

No entanto, as estruturas de conhecimento mais promissoras para expansão de consultas são as ontologias, devido ao seu alto poder de expressividade na formalização dos conhecimentos (ANDREOU, 2005; BHOGAL; MACFARLANE; SMITH, 2007; FU; JONES; ABDELMOTY, 2005; NAVIGLI; VELARDI, 2003; RAZA *et al.*, 2019; YUNZHI *et al.*, 2016). Um fator que reforça a escolha das ontologias como estruturas de conhecimento para expansão de consultas é o franco crescimento de sistemas, aplicativos e soluções “inteligentes”, decorrentes das pesquisas na área da Web Semântica, que impulsionaram significativamente o interesse pela criação e utilização das ontologias.

Dentre os vários estudos que fazem uso de ontologias, encontrados durante o levantamento bibliográfico, destacam-se nos próximos parágrafos os que mais se relacionam com a dinâmica de funcionamento do método proposto nesta tese. Ressalta-se que foram desconsiderados trabalhos que utilizaram ontologias somente na etapa de indexação.

O propósito deste levantamento foi compreender como as ontologias são utilizadas para melhorar a recuperação de informação e quais os benefícios relatados. No entanto, observou-se que apesar de resultados positivos as abordagens se mostraram complexas a parir do referencial dos utilizadores dos sistemas, sendo que algumas necessitam de construir consultas em linguagens padronizadas para consultas de grafos.

A proposta de Bonino *et al.* (2004) consiste no desenvolvimento de um sistema em que a consulta é refinada com base nas relações hierárquicas entre os conceitos da ontologia, por meio da especificação e generalização automática dos termos os autores relataram que nos experimentos realizados houveram melhorias significativas nas taxas de precisão e revocação.

Os trabalhos de Fu, Goh e Foo (2005) e Dey *et al.* (2005) baseiam-se na utilização de duas ontologias de domínio para expansão das consultas. Ambos mostraram aumentos na precisão dos resultados obtidos com as abordagens. Enquanto os primeiros trabalharam com uma ontologia para dados geográficos e outra no domínio do turismo os últimos utilizaram uma ontologia sobre vinhos e outra sobre plantas.

Para a determinação das condições de expansão Dey *et al.* (2005) calcularam a distância semântica entre os termos de consulta e os conceitos das duas ontologias. Seus experimentos consistiram na utilização do mecanismo de busca Google com as consultas expandidas, os resultados obtidos mostraram aumentos significativos de precisão.

Uma abordagem com um funcionamento um pouco diferente é a de Nilsson, Hjelm e Oxhammar (2006) que utilizaram uma ontologia de domínio baseada no sistema SUiS (*Stockholm University Information System*) para a expansão da consulta. SUiS é diferente de outros sistemas de busca pois não permite a entrada de consultas em texto livre, apenas questões como “quem, o quê, quando e onde” são permitidas. Desta forma ao invés de expandir consultas com base em todos os relacionamentos disponíveis na ontologia, apenas sinônimos e hipônimos são usados para aumentar a precisão. Os experimentos mostraram melhorias nos resultados.

As pesquisas supracitadas fazem uso das ontologias como aparatos internos nos sistemas desenvolvidos, no entanto, Ferneda (2013) ressalta que a partir de uma interface adequada as

ontologias podem servir também como ferramentas para seleção dos termos que irão compor a consulta inicial do usuário. Este tipo de abordagem permite que uma pessoa leiga em determinada área do conhecimento seja capaz de elaborar consultas pertinentes em um sistema de recuperação de informação. Além disso enquanto ela utiliza esse sistema ela adquire familiaridade com o domínio de interesse.

Explorando este tipo de abordagem Sack (2005) demonstrou como uma ontologia de domínio pode aumentar a eficiência de um sistema de recuperação de informação tradicional. Sua pesquisa utilizou uma base de dados bibliográfica e uma ontologia de problemas NP-completos. Além de ser utilizada na fase de formulação de consultas para fins de expansão e resolução de ambiguidades, essa ontologia é disponibilizada de um modo interativo, que possibilita a expansão por meio da sugestão de termos semanticamente relacionados como sinônimos, termos específicos e genéricos. O autor aponta as vantagens do uso de uma ontologia ao fornecer aos usuários um conhecimento contextualizado.

Navigli e Velardi (2003) apresentam um estudo que utiliza informações ontológicas para extrair o domínio semântico de uma palavra, identificando seu "sentido", posteriormente são feitos alguns tipos de expansões buscando melhorar os resultados das buscas. As expansões buscam resolver os problemas de sinonímia e hiperonímia além de agregarem palavras que representam o sentido encontrado. Os experimentos são feitos utilizando o mecanismo de busca do Google e a coleção de teste TREC Web 2001[6]. Os autores relatam melhorias consistentes na precisão das consultas testadas.

Neste sentido Stojanovic (2005) também observa os benefícios de apresentar ao usuário refinamentos potencialmente úteis para o aprimoramento da consulta. Em sua pesquisa o autor apresenta uma abordagem que faz a detecção do tipo de ambiguidade que a consulta pode ter utilizando ontologias e posteriormente apresenta ao usuário os termos mais adequados para à composição de sua consulta.

Refletindo sobre as pesquisas apresentadas pode-se concluir que métodos que buscam melhorar a recuperação de informação por meio do aprimoramento dos processos de elaboração das consultas são bastante importantes, pois atuam auxiliando o usuário, que por sua vez, ocupa um papel principal no processo de recuperação como um todo.

---

[6] *Text REtrieval Conference* (TREC) é um workshop anual organizado pelo Instituto Nacional de Padrões e Tecnologia do governo dos Estados Unidos, que fornece a infraestrutura necessária para avaliação em grande escala de metodologias de recuperação de texto. Uma das áreas de foco do evento é a Web, representada pela trilha Web (*Web Track*). Disponível em: https://trec.nist.gov/. Acesso em 26 mar. 2020.

Acredita-se que métodos que direcionem seus esforços em auxiliar o usuário na representação de suas necessidades informacionais são especialmente promissores, pois concentram-se na elaboração de consultas que descrevam suas reais necessidades, apesar das dificuldades inerentes ao processo.

Desconsiderar a importância do usuário, implementando sistemas e soluções com um alto grau de automatização com técnicas que aprimorem as representações internas aos SRI e que não possibilitem sua participação e controle no processo de busca de informações é privá-lo de descobertas, aprendizados e uma aquisição ativa de conhecimentos.

Desta forma, o método que está sendo proposto nesta tese consiste na contextualização e expansão de consultas baseado na utilização de ontologias de domínio como estruturas de conhecimento independentes do acervo. Sua dinâmica de funcionamento é interativa e colabora com o usuário, atuando como uma facilitadora do processo de busca e aquisição do conhecimento, que são parte primordial do objetivo desta pesquisa.

4.3 Utilização do método em um SRI

De forma sucinta, o método proposto pode ser entendido como um recurso, uma funcionalidade adicional em um mecanismo de busca cuja função é auxiliar o usuário a elaborar consultas capazes de localizar informações mais relevantes em menos tempo e com menor esforço.

Ele foi idealizado para utilização na busca de documentos textuais digitais, no entanto, não há indícios de que ele não possa ser aplicado em sistemas de recuperação de informação multimídia. No entanto, ressalta-se que este estudo se limita a discutir sua aplicação apenas em documentos textuais no ambiente Web.

Como já discutido os SRIs funcionam basicamente comparando as representações dos documentos do acervo com a expressão de busca elaborada pelo usuário, que é a representação de suas necessidades informacionais. O resultado desta comparação é uma lista de documentos ordenados pela relevância em relação à expressão de busca (BAEZA-YATES; RIBEIRO-NETO, 2013, p.23).

De forma sucinta as representações dos documentos são reduções, sumarizações de seus conteúdos que tem por objetivo torná-los encontráveis, elas são construídas por meio do processo de análise documentária, definido por Cunha (1989, p.40) como “um conjunto de

procedimentos efetuados com o fim de expressar o conteúdo de documentos, sob formas destinadas a facilitar a recuperação da informação".

Por outro lado, as expressões de busca, compostas por um conjunto de palavras escolhidas pelo usuário para traduzir suas necessidades informacionais. O processo de escolha dos termos que irão compor a expressão de busca depende de diversos fatores internos do usuário, como sua compreensão do domínio e da terminologia e também da quantidade de tempo e esforço dedicados à tarefa.

Dedicado ao auxílio do usuário na tarefa de elaboração da expressão de busca o método de contextualização e expansão de consultas tem seu funcionamento dividido em quatro passos: especificação da busca, contextualização da busca, determinação do conceito (ou classe) inicial e expansão da consulta.

4.3.1 Especificação da busca

A especificação da busca é a etapa inicial de um processo de busca por informação. Os usuários expressam suas necessidades informacionais utilizando uma ou mais palavras. A escolha do idioma das palavras assim como a estratégia de busca fica totalmente a critério do usuário, que faz estas escolhas buscando encontrar a informação desejada.

No método proposto, ao término da elaboração de sua consulta e submissão ao mecanismo de busca são disparados dois processos: a recuperação de informação e a contextualização da busca.

A recuperação de informação é o comportamento clássico dos mecanismos de busca, onde o sistema procura no acervo por documentos que contém total ou parcialmente os termos informados na consulta. O resultado é formado por uma lista de documentos supostamente contém informações que podem satisfazer as necessidades de informação do usuário.

O segundo processo diz respeito a contextualização da busca, que caracteriza a segunda etapa deste método. Seu funcionamento consiste em procurar por correspondências terminológicas nas classes das ontologias utilizando as palavras que compõem a consulta, os resultados são apresentados ao usuário em formato de uma lista.

4.3.2 Contextualização da busca

A contextualização da busca é a segunda etapa do método, sua primeira ação é a apresentação de uma listagem das ontologias que contém ou mais termos da expressão de busca. Neste contexto as ontologias representam os domínios que poderão ser utilizados para a contextualização da busca.

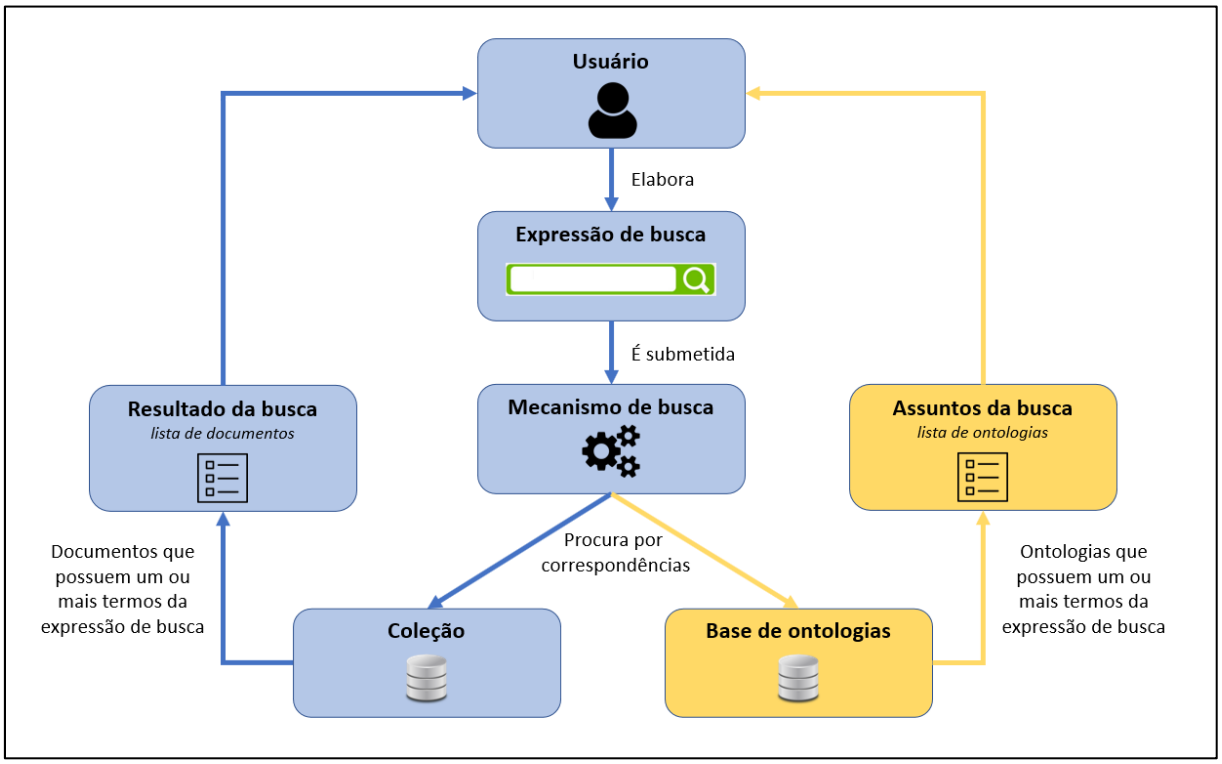
É importante ressaltar que parte-se da premissa que uma ou mais ontologias existam no SRI, sendo assim uma das limitações do método é que ele é capaz de contextualizar e expandir consultas pertencentes aos domínios das ontologias disponíveis no sistema.

Sendo assim é possível deduzir que um mecanismo de busca com o método de proposto é capaz de apresentar resultados positivos apenas nas áreas do conhecimento das ontologias existentes, nos demais domínios ele não terá efeito algum.

Considerando um exemplo hipotético em que o usuário tenha informado a expressão de busca “espinha” os domínios apresentados (provenientes das respectivas ontologias) poderiam ser: a) Doenças ou distúrbios, na qual as espinhas se referem a inflamações na pele e b) Anatomia, com a espinha se referindo às estruturas ósseas presentes nos animais vertebrados.

A Figura 15 apresenta um diagrama do funcionamento da etapa de contextualização da busca, os retângulos e setas em cor azul representam a dinâmica de funcionamento de um mecanismo de busca comum, ao passo que os componentes em amarelo representam as ações da etapa de contextualização da busca.

Figura 15 – Diagrama de funcionamento da etapa de contextualização da busca



Fonte: Desenvolvido pelo autor

O diagrama apresentado na Figura 15 começa com a ação de um usuário que elabora uma expressão de busca e a submete a um mecanismo de busca, que por sua vez procura por correspondências entre cada um dos termos da consulta em duas bases distintas: na coleção de documentos e na coleção de ontologias. Esta tarefa origina dois objetos, o conjunto de resultados da busca e uma lista dos possíveis domínios da consulta. Cada um deles é apresentado ao usuário por meio da interface do SRI.

Neste momento o usuário pode analisar os documentos nos resultados da busca e pode ainda escolher dentre os domínios apresentados qual opção melhor representa sua necessidade de informação, ação que conduz a próxima etapa: a determinação do conceito da busca.

4.3.3 Determinação do conceito (classe)

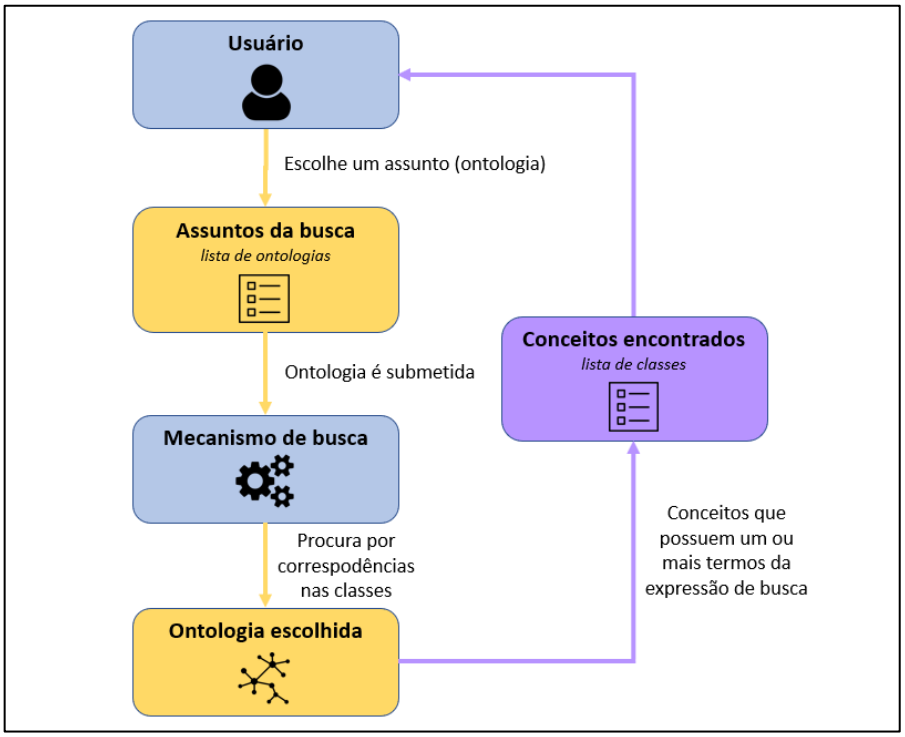
A terceira etapa do método é a determinação da classe que será o ponto de partida para a identificação de conceitos relacionados aos termos da consulta. A necessidade desta etapa origina-se da possibilidade de existência de múltiplas classes (conceitos) dentro de uma mesma ontologia que podem corresponder aos termos da expressão de busca.

Utilizando o exemplo anterior, no domínio da anatomia o termo “espinha” pode corresponder aos conceitos “Espinha dorsal” e “coluna vertebral”. Faz-se necessária então a escolha de um deles para que o mecanismo seja capaz de procurar outros conceitos utilizando as relações hierárquicas e de equivalência.

Ressalta-se que nesta etapa do processo é importante que sejam apresentados os conceitos com suas respectivas definições (ou descrições), para que o usuário tenha condições de escolher o conceito que melhor representa sua necessidade de informação.

A Figura 16 apresenta o diagrama de funcionamento da etapa de determinação do conceito. Note que há blocos em amarelo, remanescentes da etapa de contextualização e também um bloco em cor lilás que representa a etapa de determinação do conceito.

Figura 16 – Diagrama de funcionamento da etapa de determinação do conceito



Fonte: Desenvolvido pelo autor

A etapa de determinação do conceito ilustrada na Figura 16 tem início na lista dos domínios que contextualizam a busca, oriunda da etapa anterior. O domínio (ontologia) escolhido é submetido ao mecanismo de busca, que é responsável por procurar por correspondências entre os termos da expressão de busca e os rótulos e descrições das classes da ontologia escolhida.

Este processo resulta em um conjunto de conceitos (uma lista de classes) que é apresentada ao usuário na interface do sistema. A escolha de um destes itens dá início a quarta etapa do método, a identificação de conceitos relacionados.

4.3.4 Identificação de conceitos relacionados

A etapa de identificação de conceitos relacionados é feita por uma biblioteca de consulta a modelos de dados semânticos, em RDF/RDFS ou OWL, utilizando como ponto de partida a classe escolhida na etapa anterior. Por meio das relações hierárquicas e de equivalência semântica são localizadas outras classes que podem representar conceitos relacionados à expressão de busca.

Para a composição do conjunto de conceitos relacionados são exploradas as relações de hierarquia em até dois níveis de distância (exaustividade), localizando conceitos genéricos e específicos em relação à classe inicial. Além disso são exploradas as relações de equivalência entre classes, considerando todas as encontradas.

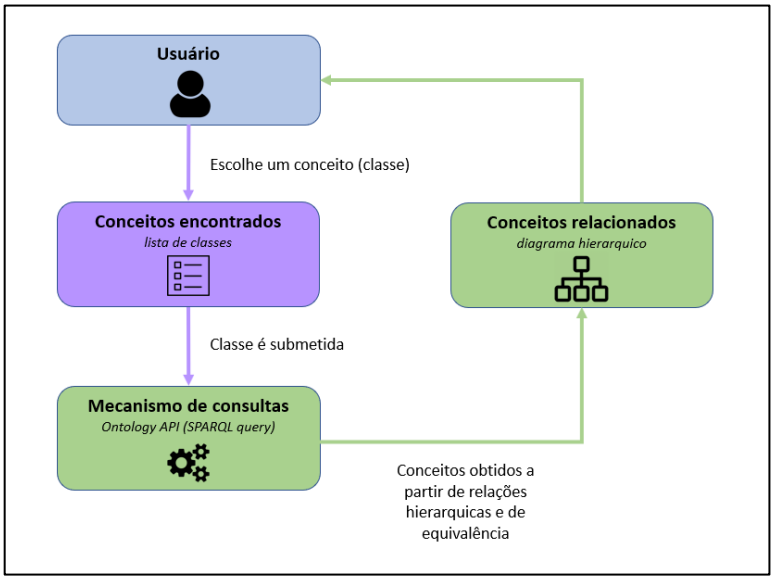
A apresentação dos resultados desta etapa é feita por meio de um diagrama da hierarquia de conceitos identificados, com destaque à classe inicial. Busca-se com isso permitir ao usuário a compreensão dos relacionamentos entre os termos do domínio de seu interesse.

Essa forma de apresentação pode propiciar ao usuário condições de especializar ou generalizar sua consulta, idealmente obtendo resultados mais específicos e precisos ou mais gerais e numerosos respectivamente.

Os conceitos equivalentes apresentados têm como objetivo a desambiguação semântica, possibilitando que o usuário agregue a expressão de busca termos mais adequados à representação de sua necessidade informacional.

A Figura 17 apresenta o diagrama de funcionamento desta etapa, os blocos em verde representam o processo de identificação de conceitos relacionados. Ela tem início com a submissão do conceito escolhido na etapa anterior ao mecanismo de consulta as ontologias, que se encarrega de procurar os termos que irão compor o diagrama hierárquico.

Figura 17 – Diagrama de funcionamento da etapa de identificação de conceitos relacionados



Fonte: Desenvolvido pelo autor

Na Figura 17 é possível notar ainda que o bloco dos “Conceitos relacionados” aponta para o usuário, isto se dá em razão do propósito desta etapa, que é apresentar ao usuário, por meio de um recurso visual, todos os termos encontrados pelo motor de inferências. Este diagrama de conceitos deve ser construído de forma a deixar claras as relações de hierarquia, superiores, inferiores e equivalentes, para que o usuário possa escolher quais termos deseja incluir em sua consulta

4.3.5 Expansão da consulta

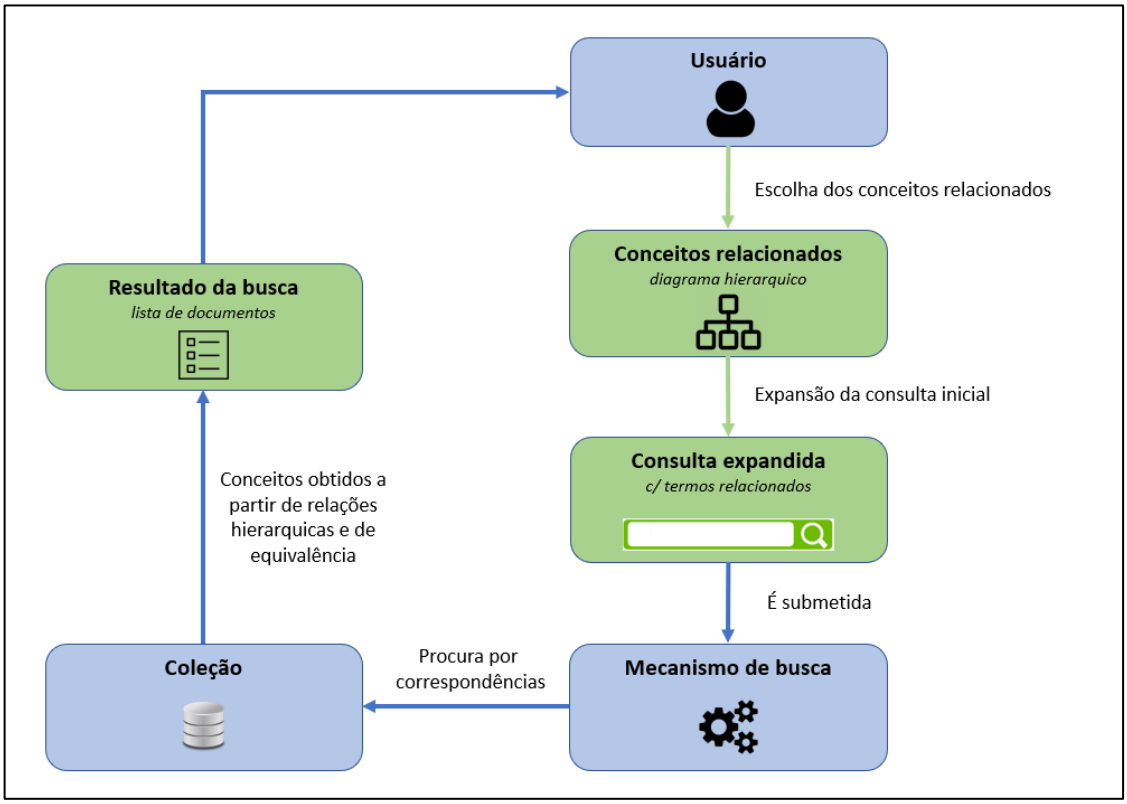
A quinta e última etapa consiste na expansão da consulta utilizando os conceitos relacionados apresentados no diagrama hierárquico. Esta etapa tem início quando o usuário opta por adicionar à sua consulta um ou mais do diagrama. O usuário pode ainda remover os termos iniciais e manter apenas os novos.

Após terminar a adição e/ou remoção de termos o usuário deve refazer a busca, submetendo novamente a expressão ao mecanismo de pesquisa, que buscará novamente no acervo por documentos que possuam coincidências terminológicas.

Um novo conjunto de resultados de busca é apresentado na interface do sistema, são esperadas duas diferenças na nova listagem de documentos: os novos registros recuperados e a uma classificação de relevância mais precisa em função de uma representação das necessidades informacionais do usuário mais adequada.

A Figura 18 apresenta o diagrama de funcionamento da etapa de expansão da consulta, com destaque aos blocos da consulta e dos resultados da busca, que estão em cor verde para demonstrar que eles sofreram alterações em relação à pesquisa inicial (anterior à utilização do método) apresentada na Figura 15.

Figura 18 – Diagrama de funcionamento da etapa de expansão da consulta



Fonte: Desenvolvido pelo autor

Após essa segunda iteração o usuário pode analisar o novo conjunto de resultados e ponderar se sua necessidade de informação foi sanada, em caso negativo pode-se iniciar uma nova iteração, retornando à consulta inicial ou ainda alterando a consulta com os termos disponíveis no diagrama hierárquico.

A utilização do método de contextualização e expansão de consulta em um SRI discutida até aqui é estritamente conceitual, no entanto, é plenamente possível implementá-la com as tecnologias existentes. No próximo capítulo é apresentado um protótipo de *software* de um mecanismo de busca Web que implementa o método proposto.

5 ContextOnSearch: um sistema de recuperação de informação com recursos de contextualização e expansão de consultas

Este capítulo descreve o processo de desenvolvimento de um protótipo de *software* que tem como objetivo demonstrar a utilização do método proposto em sistema de recuperação Web. São abordadas as tecnologias e etapas necessárias à sua consecução e por fim são apresentados os resultados obtidos a partir de uma análise qualitativa.

O método proposto nesta pesquisa tem como objetivo primário auxiliar usuários de sistemas de recuperação de informação a representar melhor suas necessidades informacionais por meio da contextualização e da expansão de consultas.

Buscando exemplificar e elucidar seu funcionamento, foi implementado um mecanismo de busca Web denominado ContextOnSearch, cuja função metodológica é demonstrar que o método é aplicável no contexto de sistemas de recuperação de informação no ambiente Web, independente da forma como foi construída a representação dos documentos que compõem o acervo.

Nesse sentido a limitação mais significativa do método é a necessidade de ontologias de domínio dotadas de relações hierárquicas e de equivalência entre seus conceitos. Sendo assim é possível utilizá-lo tanto em mecanismos de busca abertos e de propósito geral quanto em ambientes fechados e específicos.

É importante salientar que aspectos de desempenho de *software*, como performance, escalabilidade ou tolerância a falhas não foram considerados neste estudo. Ressalta-se ainda que os conceitos e requisitos da arquitetura de informações e do design de interfaces não foram observados, pois acredita-se que o desenvolvimento do protótipo não é o objetivo central da pesquisa, mas sim a discussão do método de contextualização e expansão de termos baseado em ontologias.

A partir da implementação do método em um mecanismo de busca foi possível fazer uma análise qualitativa em um ambiente controlado, que resultou em um conjunto de dados que possibilitou os resultados e análises apresentados no decorrer deste capítulo.

O ContextOnSearch utiliza em seu núcleo o ElasticSearch, um *software* de código aberto construído sobre o Apache Lucene que implementa várias funcionalidades necessárias a

um mecanismo de busca. A escolha desta tecnologia se deu em razão de sua ampla utilização em soluções que lidam com vários tipos de dados, textuais, numéricos, geoespaciais, estruturados e não estruturados (ELASTIC, 2018).

Voit *et al*. (2017) explicam que existem diversas bibliotecas para implementação mecanismos de busca em texto completo, no entanto, várias delas dependem de bancos de dados ou linguagens de programação específicas, não sendo flexíveis como o ElasticSearch. Os autores ainda apresentam uma análise comparativa de desempenho de quatro *softwares*: Sphinx, Solr, Elasticsearch e Xapian, sendo o ElasticSearch o mais adequado para implementação da solução proposta pelos autores.

Luburić e Ivanović (2016) apresentam um estudo comparativo entre o Apache Solr e o ElastiSearch para o desenvolvimento de sistemas de recuperação de informação, eles concluem que ambos são boas escolhas, no entanto, apesar de o Apache Solr ser mais comum o ElasticSearch apresenta mais simplicidade, flexibilidade e possibilidade de uma implementação modular, tornando-a mais propícia para o desenvolvimento de soluções maiores e escaláveis.

Outros autores como Amato *et al.* (2018), Shah, Willick e Mago (2018) e Kumar *et al.* (2018) também escolheram o ElasticSearch como base para implementação de seus trabalhos, reforçando a percepção de que o programa é uma solução viável e satisfatória.

Nas próximas seções abordam-se detalhes do funcionamento do ContextOnSearch, à metodologia de composição da coleção de testes, as linguagens e outras tecnologias empregadas e por fim uma apresentados dos resultados obtidos.

5.1 Apresentação do *software*

O protótipo desenvolvido nesta pesquisa, denominado ContextOnSearch, consiste em um mecanismo de busca Web que faz a contextualização e a expansão da expressão de busca formulada pelo usuário utilizando ontologias de domínio. Seu nome origina-se da junção das palavras *Context*, *Ontologies* e *Search*.

O ContextOnSearch está limitado a buscas no domínio da pediatria e saúde infantil, em razão da ontologia utilizada e do acervo, que é composto por 481 documentos provenientes de

editoriais, artigos de revisão e originais extraídos do Jornal de Pediatria[7]. O domínio escolhido justifica-se pela familiaridade com o tema e com a ontologia *Pediatric Terminology.*

Observa-se que o *software* foi executado em um ambiente controlado com apenas um domínio, porém, que ele tem potencial de ser executado com múltiplas ontologias (e domínios) em coleções de âmbito geral.

As funções de um mecanismo de busca Web são: o rastreamento de sites e outros conteúdos, sua indexação, ranqueamento de acordo com sua relevância e posteriormente a apresentação de uma lista resultados em relação à uma consulta informada por um usuário (THOMÉ, 2017).

Sendo assim o ContextOnSearch possui apenas uma destas funções, que é a de apresentação de um conjunto de resultados a partir de uma expressão de busca informada por um usuário, visto que sua coleção é estática e foi construída manualmente.

Seu diferencial reside na implementação do método de contextualização da expressão de busca e, na sugestão de termos relacionados para expansão da consulta, aprimorando assim a etapa de construção de uma expressão que traduza da melhor forma possível as necessidades de informação de um usuário.

A Figura 19 apresenta a interface de busca do ContextOnSearch com a expressão de busca “catapora” e seus respectivos resultados ordenados pela relevância. Nela pode-se ver quatro áreas em destaque: 1) área de busca, 2) área de resultados, 3) área de contextualização e 4) área de apresentação dos termos relacionados.

A área da busca consiste em uma caixa de texto que o usuário pode elaborar sua expressão de busca utilizando palavras ou expressões compostas em linguagem natural, assim como buscadores como Google ou Bing. Houve uma preocupação em manter essa simplicidade, pois abordagens que necessitam de passos adicionais, interações por parte dos usuários podem gerar rejeição e/ou dificuldades indesejadas.

[7] O Jornal de Pediatria é um periódico mantido pela Sociedade Brasileira de Pediatria, que publica artigos originais e de revisão abrangendo diversas áreas da pediatria. Disponível em: https://jped.elsevier.es/pt. Acesso em: 15 mar. 2021.

Figura 19 – Página de busca do ContextOnSearch

Fonte: Desenvolvido pelo autor

Na área de resultados é apresentada uma lista com os documentos que contém um ou mais termos da expressão de busca ordenados por relevância. O método de cálculo da relevância é discutido detalhadamente na seção sobre o ElasticSearch. Cada registro recuperado é composto pelo título, URL de acesso e conteúdo completo, sendo que ao clicar sobre o título o usuário é direcionado ao documento original na Web.

A área número três, denominada área de contextualização, é destinada à primeira das ações do método: a contextualização da expressão de busca, que consiste essencialmente em buscar em um acervo de ontologias por coincidências terminológicas entre as palavras que compõem a consulta e as presentes nos rótulos e comentários das classes das ontologias.

O resultado desta ação é uma lista com os possíveis domínios (contextos) da expressão de busca, composto por uma ou mais ontologias que contém classes com as palavras coincidentes. Neste momento o usuário pode escolher (ou não) um dos domínios que representa a sua necessidade de informação.

Por fim, na área número quatro é apresentada uma hierarquia de termos relacionados aos termos da consulta, obtidos por meio das relações semânticas presentes nas ontologias. Esta

ação depende diretamente da contextualização da consulta feita na área número três, que na prática permite que o usuário escolha uma ontologia que melhor representa o domínio de conhecimento de sua busca.

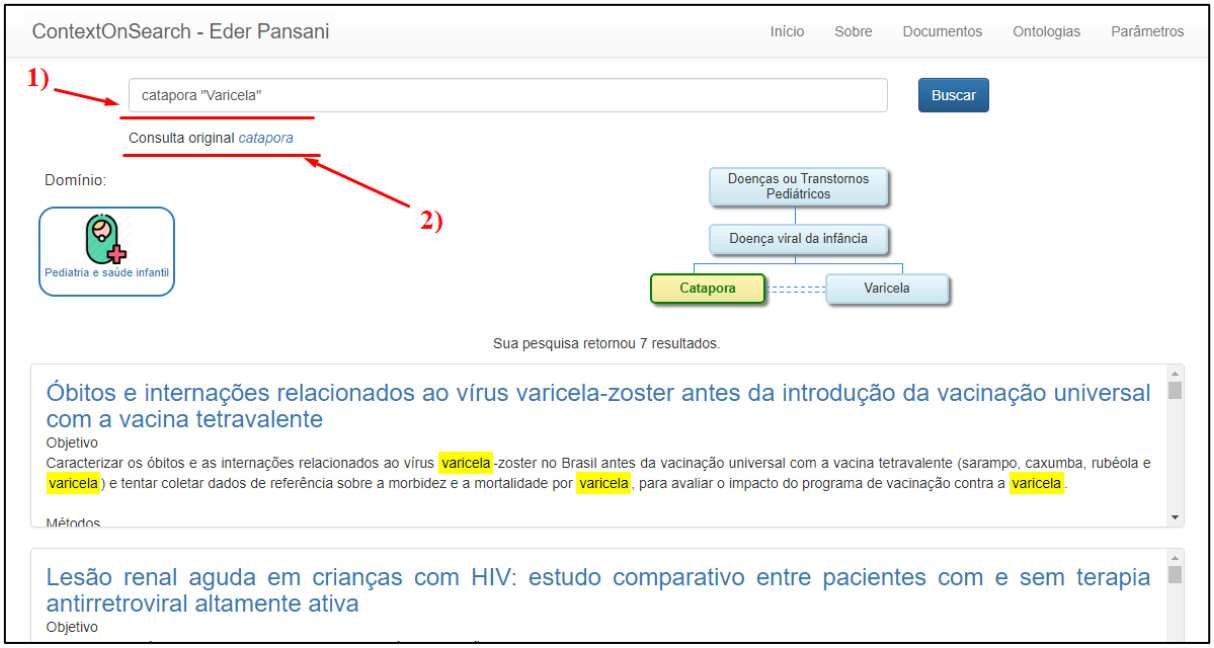
Esta hierarquia de termos relacionados permite que seja feita a expansão da consulta inicial, pois clicando sobre os termos o usuário pode adicioná-los à sua expressão de busca e refazer a sua busca com intento de se obter resultados mais relevantes.

Ressalta-se que este processo de expansão da consulta é inteiramente manual e dependente das decisões do usuário durante a utilização do sistema.

Na Figura 20 pode-se observar o funcionamento do ContextOnSearch com a expansão de consultas. Na área destacada número 1 fica a caixa de texto com a consulta expandida, os termos que foram adicionados ficam entre aspas para diferenciá-los dos termos que compõem a consulta inicial.

Na segunda área destacada, logo abaixo da caixa de texto, fica a consulta original em formato de *hiperlink*, permitindo que seja possível retornar à busca inicial, caso o usuário julgue que os resultados não são pertinentes.

Figura 20 – Página de busca com destaque as consultas expandida e original



Fonte: Desenvolvido pelo autor

O ContextOnSearch possui ainda algumas funcionalidades secundárias, tais como:

- gerenciamento de documentos: composto por um formulário de inserção de documentos na coleção e por uma página de listagem;

- gerenciamento de ontologias: composto por formulário de importação de ontologias do Apache Jena e adição do rótulo e do ícone que serão mostrados na área de contextualização e por uma página de listagem das ontologias disponíveis no sistema;
- parâmetros de sistema: página de configuração de alguns parâmetros de sistema, como a apresentação de destaque nas palavras nos documentos entre outros.

Os resultados da execução de algumas buscas utilizando o ContextOnSearch são apresentados em uma seção dedicada. No entanto, antes de prosseguir para essa seção faz-se necessário discutir as questões da composição da coleção de documentos e de ontologias que foram utilizados na implementação.

5.2 Documentos: a composição do *corpus*

O método proposto neste estudo foi pensado para utilização em mecanismos de busca Web, independente de assunto ou acervo de documentos. No entanto, para seu funcionamento é necessário que haja uma ou mais ontologias de domínio que tratem dos assuntos dos documentos pertencentes ao corpus.

Desta forma, o método proposto pode ser aplicado a qualquer mecanismo de busca, porém, os processos de contextualização e a expansão de consultas serão feitos apenas para os domínios de conhecimento das ontologias disponíveis no sistema.

Para superar essa barreira optou-se por trabalhar com um objeto delimitado, um conjunto de documentos sobre o domínio da ontologia escolhida: pediatria e saúde infantil. Para obtenção destes documentos foi feita uma coleta das publicações no Jornal de Pediatria, um periódico mantido pela Sociedade Brasileira de Pediatria.

De acordo com seu *website*, o Jornal de Pediatria é

> uma publicação bimestral da Sociedade Brasileira de Pediatria, em circulação desde 1934, o periódico publica artigos originais e artigos de revisão, abrangendo as diversas áreas da pediatria. Através da publicação e divulgação de relevantes contribuições científicas da comunidade médico-científica nacional e internacional da área de pediatria, o Jornal de Pediatria busca elevar o padrão da prática pediátrica e do atendimento médico especializado em crianças e adolescentes.

Devido à periodicidade bimestral, em cada ano é publicado um novo volume composto por seis números, 1 – janeiro e fevereiro, 2 – março e abril, 3 – maio e junho e assim

sucessivamente. Foram coletados todos os trabalhos publicados nos últimos 5 anos (2016 a 2020) em cada um dos seis números de cada volume, totalizando 481 registros.

O processo de coleta dos documentos foi realizado entre os dias 2 e 8 de dezembro de 2020, e consistiu em acessar cada um dos trabalhos e copiar manualmente as informações para um formulário no ContextOnSearch.

O formulário de inserção de documentos na coleção conta com os campos ano, volume, número, título, URL e conteúdo, e pode ser visto na Figura 21.

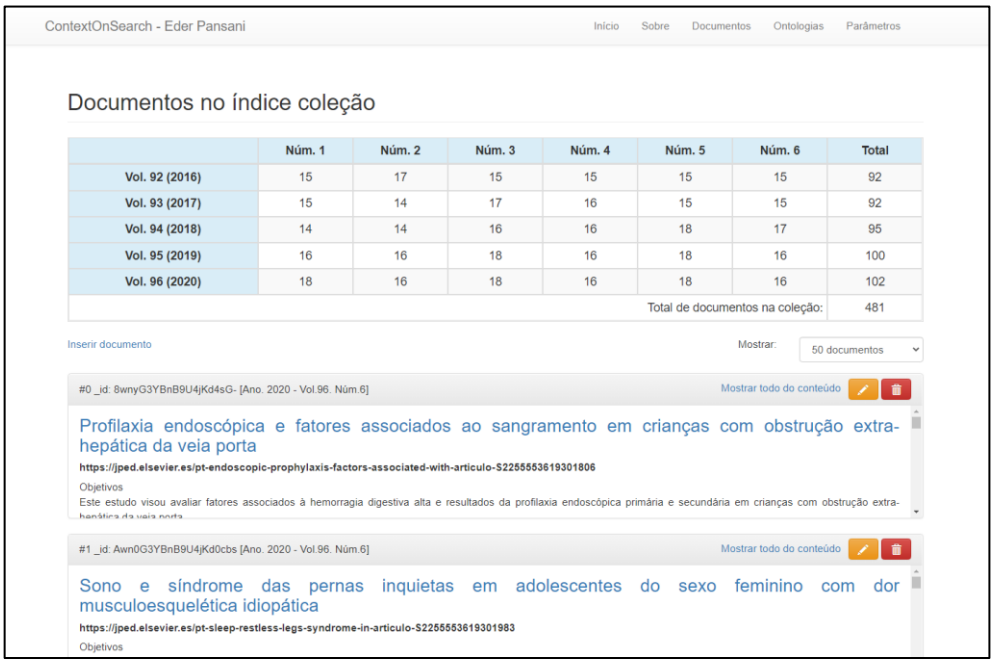
Figura 21 – Formulário para inserção de documentos na coleção

Fonte: Desenvolvido pelo autor

A separação do título e do conteúdo se deu para uma melhor apresentação dos documentos nas listagens na interface da aplicação. Já na URL possibilita a criação de um hiperlink para acessar o registro original no sítio do Jornal da Pediatria e os demais campos foram armazenados para fins de organização interna.

A Figura 22 apresenta a página de listagem dos documentos na coleção. No início há uma tabela com a quantidade de trabalhos por volume e número e na sequência observa-se a listagem dos registros.

Figura 22 – Listagem de documentos na coleção



Fonte: Desenvolvido pelo autor

A delimitação deste objeto finito, como acervo do protótipo desenvolvido (ContextOnSearch), foi importante para que fossem possíveis algumas análises e simulações de resultados de buscas em um experimento controlado. Os resultados das buscas com os termos escolhidos são descritos na seção de resultados obtidos.

5.3 Ontologias

Como já discutido, o método proposto necessita de um conjunto de ontologias de domínio que tenham um número significativo de classes, que por sua vez possuam descrições de seus conceitos e relações semânticas de hierarquia e equivalência.

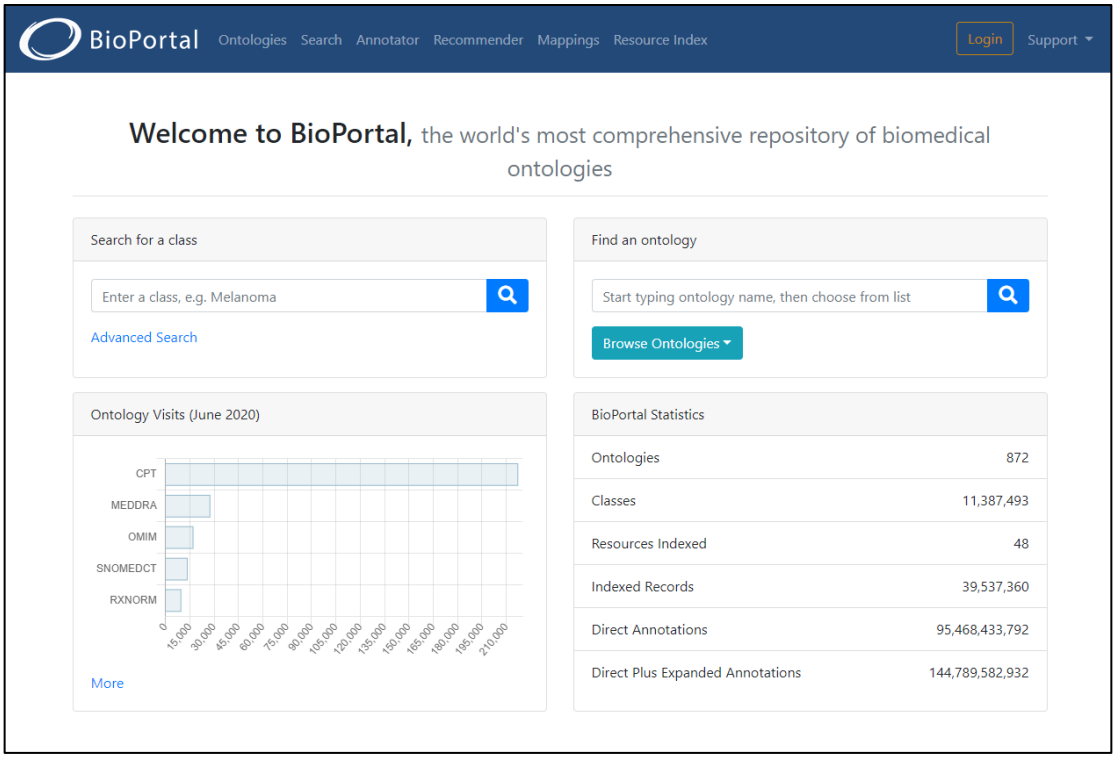
A ontologia escolhida para utilização no ContextOnSearch foi a *Pediatric Terminology* (PEDTERM), uma ontologia com "Termos associados à pediatria, representando informações relacionadas à saúde e desenvolvimento infantil desde o pré-natal até os 21 anos de idade [...]" (BIOPORTAL, Não paginado, 2013, tradução nossa), desenvolvida com a colaboração do *National Institute of Child Health and Human Development* (NICHD).

A ontologia *Pediatric Terminology* está disponível no BioPortal[8], um repositório de ontologias na área biomédica, que se declara como "[...] o repositório mais abrangente do

[8] Disponível em: https://bioportal.bioontology.org/. Acesso em 15 mar. 2021.

mundo de ontologias biomédicas.” (BIOPORTAL, Não paginado, 2021, tradução nossa). Na Figura 23 pode-se ver a página inicial do repositório.

Figura 23 – Página inicial do repositório de ontologias BioPortal



Fonte: Repositório de ontologias BioPortal (2020)

Dentre as razões para escolha desta ontologia encontram-se:

- a familiaridade com a ontologia e com o assunto, pois a mesma já foi utilizada em outras pesquisas (PANSANI-JUNIOR, 2016);
- o formato, a ontologia PEDTERM é construída usando a linguagem OWL atendendo aos requisitos do projeto;
- o conteúdo, esta ontologia possui uma grande quantidade de conceitos interrelacionados em uma hierarquia de classes além de definições e sinônimos declarados, o que a torna bastante rica do ponto de vista da representação do domínio de conhecimento.

Contando com 1771 classes, a PEDTERM teve seu desenvolvimento iniciado em 2011 e uma segunda versão em 2013, sendo esta sua última atualização e a versão utilizada neste trabalho, que pode ser obtida também em formatos CSV e RDF/XML.

A ontologia PEDTERM foi criada originalmente no idioma inglês com suas classes e propriedades nomeadas utilizando o estilo híbrido *CamelCase* com *Underscore*, como a classe

"*Disease_or_Disorder*" por exemplo. Um fato importante sobre a PEDTERM é que ela não utiliza rótulos (*tag rdfs:label*) para nomear suas classes, o que não significa que ela está "incorreta", visto que de acordo com a documentação da linguagem OWL a utilização de rótulos é facultativa.

No entanto, para o método proposto isso configura um problema, pois o mecanismo de inferência faz uso deles para comparar com os termos da consulta do usuário. Sendo assim, foi necessário fazer ajustes na ontologia, criando os rótulos em algumas classes para possibilitar a execução das buscas.

A Figura 24 mostra um trecho da ontologia com a classe "*Disease_or_Disorder*" alterada, no trecho destacado (linhas 28 a 32) estão os rótulos adicionados. Optou-se ainda pela criação dos rótulos no idioma original (inglês) e também em português e espanhol para uma abordagem multilíngue.

Figura 24 – Trecho da ontologia PedTerm, exemplo de criação dos rótulos (*labels*)

```
15  <owl:Class rdf:ID="Disease_or_Disorder">
16      <ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
17      >C2991</ID>
18      <NCI_PT rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
19      >Disease or Disorder</NCI_PT>
20      <NCI_Definition rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
21      >Any abnormal condition of the body or mind that causes discomfort, dysfunction, or d
22  <SYN rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
23      >Disease</SYN>
24      <Subset_Association1 rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
25      >NICHD Pediatric Terminology</Subset_Association1>
26  <Subset_Association2 rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
27      >Neonatal Research Network Terminology</Subset_Association2>
28      <!-- Alteração Eder Pansani -->
29        <rdfs:label xml:lang="en">Disease or Disorder</rdfs:label>
30        <rdfs:label xml:lang="es">Emfermedad o Transtorno</rdfs:label>
31        <rdfs:label xml:lang="pt">Doença ou Disturbio</rdfs:label>
32      <!-- Alteração Eder Pansani -->
33  </owl:Class>
```

Fonte: Desenvolvido pelo autor com base na ontologia PEDTERM

Ressalta-se que a criação dos rótulos foi feita apenas em algumas classes, possibilitando a execução dos testes apresentados na seção de resultados obtidos. Acredita-se que alterar todas as classes pertencentes a esta ontologia não seria uma tarefa pertinente aos propósitos da pesquisa.

O passo seguinte foi carregar a ontologia no Apache Jena Fuseki, para tornar seu conteúdo acessível utilizando a linguagem SPARQL. O ContextOnSearch se comunica diretamente com o Apache Jena Fuseki por meio de um *Endpoint Sparql*. No entanto, para o pleno funcionamento do sistema é necessário fazer a importação de todas as classes das ontologias para um índice no ElasticSearch.

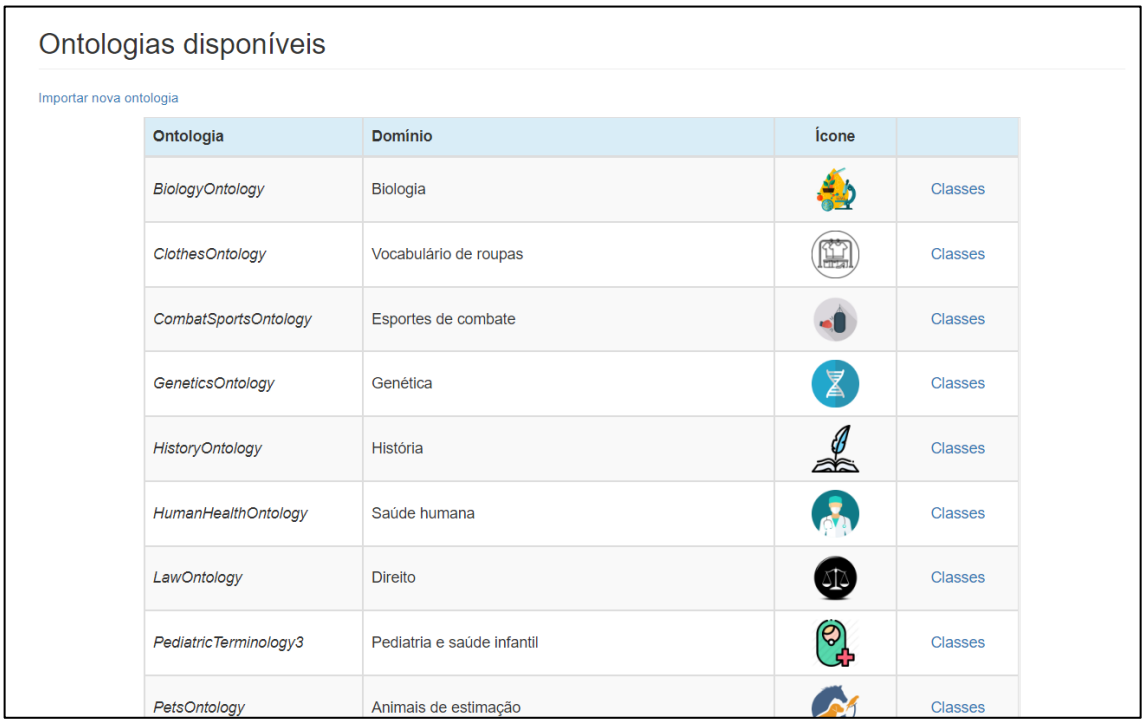
Este processo é feito em um formulário desenvolvido para importação de ontologias, que pode ser visto na Figura 25. Há ainda dois outros campos neste formulário: o domínio e o ícone que estão associados a esta ontologia, informações que são utilizadas na área de contextualização da expressão de busca.

Figura 25 – Formulário para importação de ontologias

Fonte: Desenvolvido pelo autor

As ontologias que foram importadas e estão disponíveis no sistema podem ser consultadas na página mostrada na Figura 26. Desta lista somente a "*PediatricTerminology3*" é uma ontologia verdadeira, as demais são fictícias, criadas com a finalidade de demonstrar o comportamento do ContextOnSearch quando uma determinada consulta pode pertencer a diferentes contextos.

Figura 26 – Formulário para importação de ontologias



Fonte: Desenvolvido pelo autor

Há ainda uma página que apresenta uma listagem de todas as classes de uma determinada ontologia, que pode ser acessada no *hiperlink* "Classes" em cada uma das linhas da tabela apresentada na Figura 26.

A necessidade de importação das ontologias para o ElasticSearch se justifica pela natureza do processo de contextualização, que consiste na busca pelos termos que compõem a consulta em um processo sintático feito utilizando as bibliotecas do ElasticSearch para processamento de textos. Sendo assim, não é necessário a exploração dos relacionamentos semânticos entre as classes, o que explica a importação apenas dos rótulos e comentários.

Já na etapa de identificação de termos relacionados para expansão da consulta, parte-se de um conceito escolhido pelo usuário para buscar, por meio dos relacionamentos hierárquicos e de equivalência, outros termos que podem ser agregados á consulta inicial, com objetivo de torná-la mais adequada para obtenção da informação desejada.

Para a execução deste processo, que explora os relacionamentos de um determinado conceito é necessária uma linguagem de consulta e uma biblioteca de manipulação de ontologias, que no caso do ContextOnSearch são respectivamente a linguagem SPARQL e o Apache Jena Fuseki, descritos na próxima seção.

5.4 Tecnologias utilizadas

Existem hoje centenas de tecnologias, linguagens e bibliotecas de *software* que podem ser utilizadas para o desenvolvimento de aplicações, além das novidades que surgem diariamente no campo da tecnologia da informação.

Além disso é fato que vários conjuntos de linguagens e bibliotecas podem sem empregados para obtenção de uma mesma solução de *software*, tornando a escolha a escolha de uma tecnologia uma decisão complicada.

Desta forma para a criação do ContextOnSearch foram definidas três regras que as tecnologias, linguagens e bibliotecas deveriam respeitar: ser livres (gratuitas), amplamente utilizadas e flexíveis.

A adoção de tecnologias livres se em razão da crença de que elas são essenciais para o desenvolvimento social e tecnológico de uma sociedade, pois permitem o uso, distribuição e aperfeiçoamento dos *softwares* desenvolvidos.

Já a escolha por tecnologias amplamente utilizadas se deu em razão da disponibilidade de documentação e material didático, da maturidade dos *softwares* e da facilidade de obtenção.

Em relação a flexibilidade o principal requisito foi que as tecnologias precisariam ser adequadas tanto para o desenvolvimento de soluções de grande porte, que necessitem de lidar com grandes volumes de dados, quanto para soluções menores e mais específicas, porém com problemas em comum.

Os próximos tópicos abordam sucintamente as ferramentas e linguagens de programação empregados no desenvolvimento do protótipo, com exceção do ElasticSearch, que possui uma descrição mais detalhada em razão de sua importância como elemento central da aplicação e também sua complexidade e volume de definições.

5.4.1 ElasticSearch

O ElasticSearch é responsável por prover o núcleo do OntosSearch, pois é ele quem implementa a maior parte das funcionalidades necessárias a um motor de busca. Ações como a indexação dos documentos, geração dos índices, disponibilização de uma função de busca e o ranqueamento dos resultados são implementados pela ferramenta.

O ElasticSearch é um mecanismo de busca de código aberto construído sobre o Apache Lucene, sendo assim ele é basicamente uma camada de *software* com todas as funcionalidades do Apache Lucene e com novos recursos o suporte ao processamento distribuído, que é importantíssimo considerando o volume de informações gerado pela sociedade contemporânea (GORMLEY; TONG, 2015, p.3, tradução nossa).

Operando sob licença *open source*, o ElasticSearch pode ser utilizado por qualquer pessoa de forma gratuita. Este licenciamento se aplica a quaisquer tipos de dados, desde informações textuais até dados não estruturados e registros de logs.

O nome ElasticSearch é uma marca comercial da empresa holandesa Elastic NV, que oferece serviços de planejamento, implementação e suporte para aplicações de busca, controle de logs, métricas de sistema, monitoramento e controle de disponibilidade, busca integrada em âmbito empresarial, exploração de dados de localização em tempo real entre outras (ELASTIC NV, 2018).

Dentre as razões para escolha da ferramenta destacam-se a rapidez na busca de texto completo, a flexibilidade de poder utilizá-lo em um computador pessoal ou em grande número de servidores de forma distribuída e o amplo conjunto de recursos e funcionalidades disponíveis no pacote padrão do programa (ELASTIC, 2019).

Em relação ao tempo de vida e à maturidade, acredita-se que o ElasticSearch pode ser considerado uma tecnologia consolidada, pois foi lançado em 2010 e vem recebendo aprimoramentos por parte da colaboração da comunidade e de incentivos da empresa Elastic NV, sendo utilizado por diversas organizações ao longo do tempo.

5.4.2 Apache Jena Fuseki

O Apache Jena é um *framework* Java de código aberto e utilização livre indicado para construção de aplicações de Web Semântica e *linked data* (APACHE..., [201-]). Dentre suas funcionalidades destacam-se:

- API RDF: utilizada para criar e ler grafos em RDF, possibilitando ainda a serialização de triplas em formatos RDF, XML ou *Turtle*;
- ARQ (SPARQL): fornece um mecanismo de consulta a dados em RDF compatíveis com SPARQL 1.1. Suporta ainda consultas remotas e pesquisa de texto livre;

- TDB: utilizado para persistência de dados usando o TDB, uma forma nativa de armazenamento de triplas;
- Fuseki: implementa um *Endpoint* SPARQL acessível via protocolo HTTP, possibilita interações com os dados armazenados via formato REST;
- *Ontology* API: possibilita o uso de OWL para adicionar semântica extra aos dados em formato RDF (padrão do Apache Jena);
- *Inference* API: fornece um *reasoner* (mecanismo de raciocínio) sobre os dados, possibilitando inferência e verificação dos dados usando regras próprias em OWL ou RDFS.

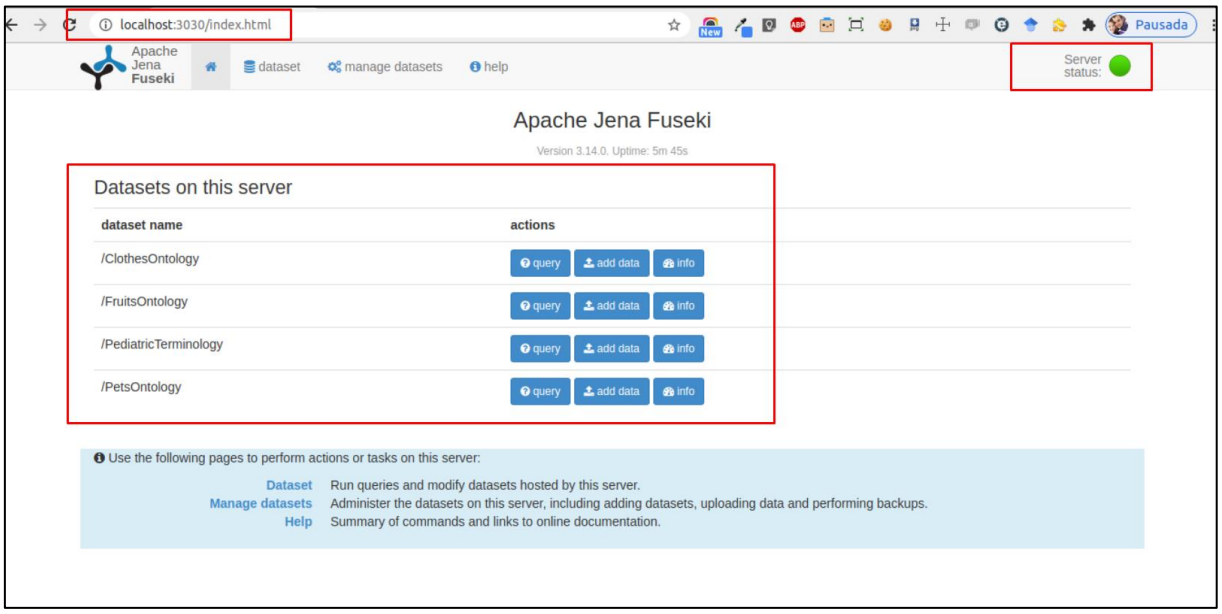
No ContextOnSearch foram utilizados os serviços *Ontology* API e Fuseki, respectivamente para o carregamento dos dados (ontologias em formato OWL) para o servidor e para a interação do *software* por meio de consultas SPARQL submetidas via requisições HTTP para o servidor do Apache Jena.

No website oficial do Apache Jena[9] há links para download do programa, tutoriais para instalação e configuração e a documentação de cada um dos serviços supracitados. Após instalado, o programa apresenta uma interface de administração que pode ser vista na Figura 27, que tem três áreas destacadas: a URL de acesso, o status e os conjuntos de dados no servidor.

A área dos conjuntos de dados (*datasets*) apresenta uma lista com os conjuntos criados a partir de ontologias enviadas ao servidor, onde, à direita de cada um deles há três opções, para executar consultas ao *dataset*, adicionar mais dados ou consultar suas informações de volume de dados e estatísticas de acesso.

[9] Disponível em: https://jena.apache.org/index.html. Acesso em 26 mar. 2020.

Figura 27 – Página inicial do servidor Apache Jena



Fonte: Desenvolvido pelo autor

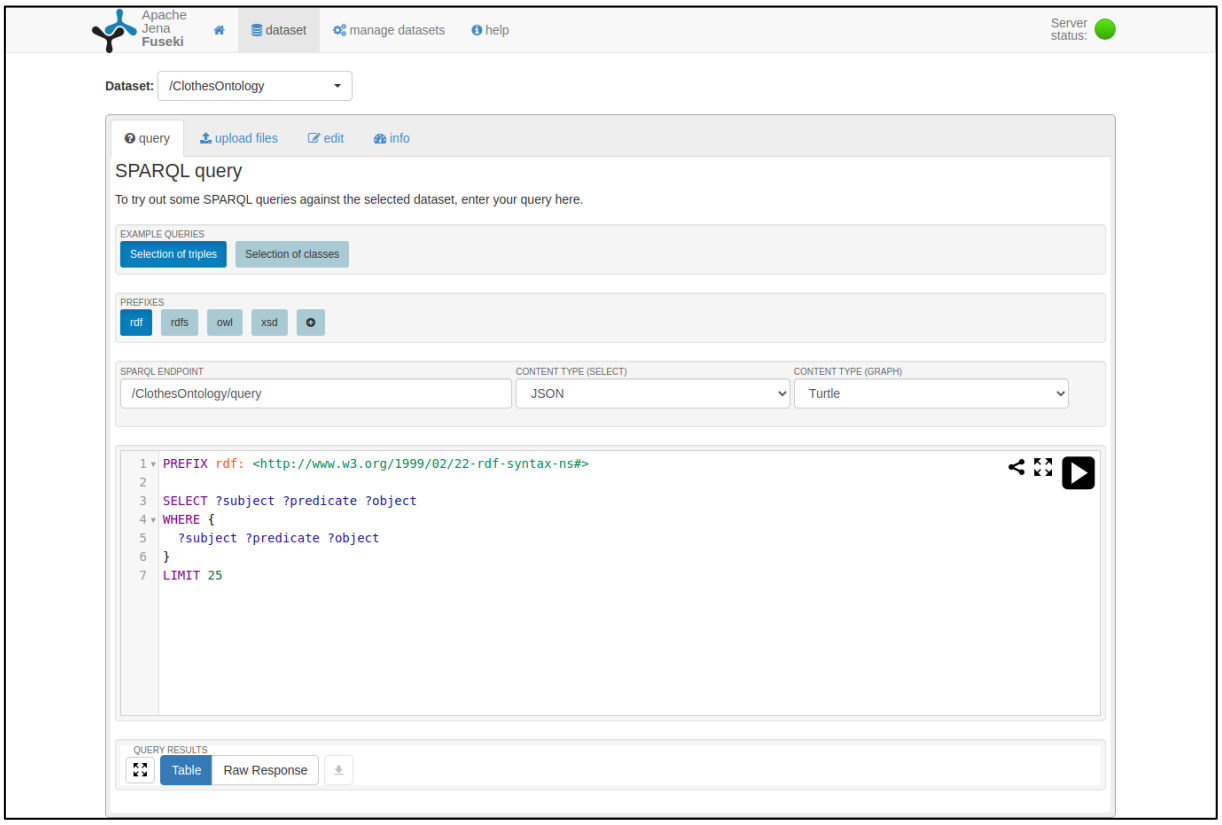
No menu superior do Apache Jena Fuseki pode-se ver três opções: "*dataset*", *manage datasets*" e "*help*". A opção "*manage datasets*" permite o gerenciamento dos conjuntos de dados, exclusão, backup e adição de novos dados. Na aba "*help*" há links para a documentação da versão instalada e, por fim, a opção "*dataset*" apresenta uma tela que possibilita a execução de consultas SPARQL, demonstrada na Figura 28.

Esta interface de execução de consultas SPARQL a um conjunto de dados é útil para o aprendizado e familiarização com a linguagem, pois há opções para adição dos prefixos ("*namespaces*") mais comuns e também para apresentação dos dados recuperados em diversos formatos, como JSON, XML ou *Turtle.*

Apesar desta interface ser amigável, ela não é utilizada pelo ContextOnSearch, que faz uso de requisições HTTP por meio da linguagem de programação PHP com a biblioteca "sparqllib.php" para consultar os conjuntos de dados e retornar as respostas para serem tratadas e apresentadas em sua interface.

A próxima subseção aborda em mais detalhes a linguagem de consultas SPARQL, que possibilitou a extração de informações armazenadas nos conjuntos de dados oriundos das ontologias de domínio.

Figura 28 – Página de consultas SPARQL ao conjunto de dados



Fonte: Desenvolvido pelo autor

5.4.3 SPARQL Query Language

O termo SPARQL, cuja pronúncia é "*sparkle*", é um acrônimo recursivo para *SPARQL Protocol and RDF Query Language*, uma linguagem usada para consultar repositórios de dados em RDF[10]. A primeira versão dessa linguagem foi lançada em janeiro de 2008. Posteriormente, em 2013 foi lançada a versão 1.1, que apresenta diversas mudanças e uma maior robustez (PRUD'HOMMEAUX; SEABORNE, 2008; HARRIS; SEABORNE, 2013).

Hortêncio-Filho e Lóscio (2009) afirmam que a linguagem é recomendada pelo W3C para a recuperação de informação em documentos no formato RDF/RDFS, tal como a linguagem SQL é usada para recuperar registros em banco de dados relacionais. Embora possam inicialmente ser parecidas, as linguagens possuem diferenças importantes, já que o SPARQL busca correspondências a partir de grafos, enquanto o SQL busca dados relacionais em formato de tabelas (WATSON, 2010).

[10] RDF é um formato de dados rotulado em grafos para representação de dados na Web frequentemente utilizado para representar metadados sobre artefatos digitais e fornecer um meio de integração entre fontes díspares de informação. Disponível em: https://www.w3.org/RDF/. Acesso em 19 ago. 2020.

DuCharme (2013, p.21, tradução nossa) explica que uma consulta SPARQL normalmente diz "[...] quero essas informações a partir do subconjunto de dados que atendem essas condições [...]", no qual você deve descrever as condições no padrão de triplas, como as triplas RDF, podendo incluir variáveis para adicionar flexibilidade na forma como eles correspondem aos dados.

O propósito do SPARQL é explorar o conhecimento representado em OWL e/ou RDF/RDFS analisando propriedades, restrições e definições que permitam a execução de inferências lógicas sobre os dados.

É importante ressaltar que ontologias OWL, como a PEDTERM utilizada neste estudo, são estruturadas utilizando OWL e RDF/RDFS para a representação adequada de um domínio do conhecimento. Cada uma das linguagens tem um propósito específico, definido por Bulcao-Neto, Prazeres e Pimentel (2006):

- o RDF define um modelo abstrato para fazer declarações sobre recursos;
- o RDFS adiciona características e fornece um vocabulário padronizado para descrever os tipos das classes, suas instâncias, restrições e relacionamentos hierárquicos;
- a OWL estende o vocabulário de esquemas RDF com construtores mais ricos em relação à expressividade e à inferência, utilizando o modelo de dados RDF.

Ducharme (2013, p.110, tradução nossa) afirma que existem quatro formas de realizar uma consulta em SPARQL:

- SELECT: é a forma principal de consultar uma coleção, os dados recuperados podem ser apresentados como uma tabela de linhas e colunas ou ainda em outro padrão definido;
- CONSTRUCT: tem a função de retornar triplas. É mais utilizado quando há necessidade de se construir novos dados a partir de combinações, conversões e replicações de dados anteriores;
- ASK: o processador de consultas verifica se um determinado padrão de grafos descreve um conjunto de triplas RDF de uma base de dados específica, retornando operadores booleanos, de verdadeiro ou falso;
- DESCRIBE: tem a função de identificar e descrever todas as triplas relacionadas a um objeto particular, especificado na consulta.

Pérez, Arenas e Gutierrez (2009) complementam explicando que uma consulta SPARQL consiste em três partes:

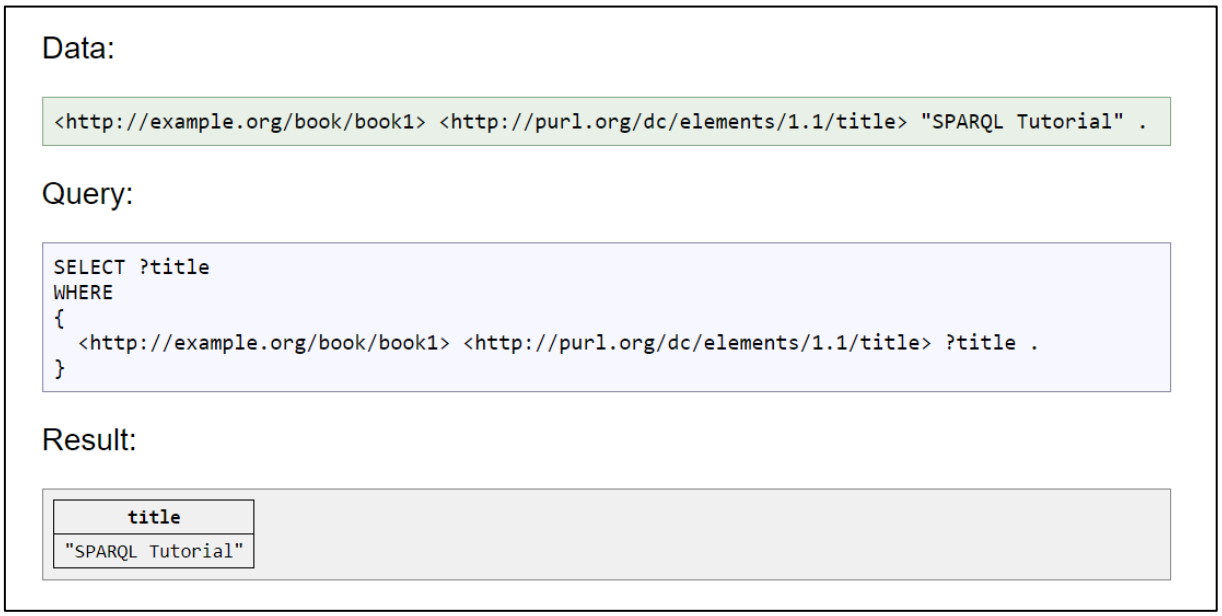
- *Pattern Maching*: inclui várias características interessantes de combinação de padrões de gráficos, a união de padrões, aninhamento, filtragem e a possibilidade de escolher a fonte dos dados;
- *Solution modifiers*: permite modificar os valores de saída padrão aplicando alguns operadores como PROJECTION, LIMIT, DISTINCT, ORDER e OFFSET;
- *Output*: podem ser de diferentes tipos, como: booleanas, seleções, gráficos e descrições de recursos.

Sendo que as três partes não são obrigatórias para o funcionamento de toda e qualquer consulta, sendo facultado seu uso dependendo das necessidades de cada situação. Observa-se ainda que no protótipo desenvolvido utiliza-se apenas a cláusula SELECT para consultar os dados e recuperá-los em formato de texto sem formatação.

O exemplo de uma consulta SPARQL simples, apresentado na Figura 29, mostra como encontrar o título de um livro a partir de dados em grafo. A consulta é dividida em duas partes: a cláusula "*SELECT*" que identifica as variáveis a serem exibidas nos resultados, no caso há somente uma variável, "*title*" e a cláusula "*WHERE*" que fornece o padrão para correspondência dos dados.

A consulta mostrada neste exemplo é bastante simples e limita-se a recuperar uma parte dos dados de uma tripla, no entanto, o SPARQL tem potencial recuperar dados de uma forma que linguagens estruturadas com o SQL não tem. Utilizando união de padrões, filtros sobre os dados, relações hierárquicas e propriedades RDF/RDFS ou OWL podem ser construídas consultas significativamente mais complexas e capazes de suprir as necessidades de informações para soluções inteligentes.

Figura 29 – Exemplo de uma consulta SPARQL simples



Fonte: Adaptado de SPARQL WORKING GROUP (2008)

Dentre as propriedades que exploram a semântica dos dados e podem ser utilizadas no âmbito da recuperação de informação destacam-se, a partir de Coneglian (2017, p.82-83), as seguintes:

- *rdfs:range e rdfs:domain*: propriedades restritivas que podem caracterizar relações de aproximação entre duas ou mais classes;
- *rdfs:subClassOf*: propriedade hierárquica que pode caracterizar uma especificação ou uma generalização de uma determinada classe;
- *owl:oneOf:* propriedade enumerativa que pode promover a ligação entre diversas classes, indicando que uma determinada classe faz parte de um conjunto de classes relacionadas;
- *owl:equivalentClass e owl:sameAs*: propriedades de equivalência que podem caracterizar classes com mesmo significado semântico;
- *owl:intersection*: propriedade restritiva que pode caracterizar classes oriundas de uma mesma classe, tendo assim alguns aspectos semânticos em comum.

Existem ainda uma ampla gama de outras propriedades, no entanto elas não são abordadas pois acredita-se que uma revisão extensiva das características das linguagens OWL, RDF e RDFS foge ao escopo desta pesquisa.

As consultas elaboradas para o ContextOnSearch limitam-se a utilizar apenas as propriedades hierárquicas e de equivalência, com objetivo de encontrar termos (conceitos) genéricos, específicos ou equivalentes em relação aos termos que compõe a expressão de busca.

Isto não significa que outras propriedades e abordagens não possam ser implementadas e agregadas ao método, maximizando seu potencial de expansão das consultas para melhor representar as necessidades informacionais dos usuários. Acredita-se que esta seja um nicho promissor para futuras pesquisas.

No ContextOnSearch são utilizadas consultas independentes para a busca por termos hierarquicamente superordenados (genéricos) e subordinados (específicos) a um determinado termo. Estas consultas podem ser vistas respectivamente na Figura 30. Observe que elas são praticamente idênticas, sendo suas diferenças indicadas pelas setas em vermelho.

Figura 30 – Consultas SPARQL utilizadas para explorar relações hierárquicas

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  PREFIX owl: <http://www.w3.org/2002/07/owl#>
3
4  SELECT (str(?label) as ?Termos)
5
6  WHERE
7  {
8    {
9      ?res1 rdfs:label ?o .
10     ?res1 rdfs:subClassOf ?res2 .
11     ?res2 rdfs:label ?label
12     FILTER(lang(?label) = "pt") .
13     FILTER(lcase(?o) = lcase("Doença Bacteriana"@pt)) .
14   }
15   UNION
16   {
17     ?res1 rdfs:label ?o .
18     ?res1 rdfs:subClassOf ?res2 .
19     ?res2 rdfs:subClassOf ?res3 .
20     ?res3 rdfs:label ?label
21     FILTER(lang(?label) = "pt") .
22     FILTER(lcase(?o) = lcase("Doença Bacteriana"@pt)) .
23   }
24 }
25
```

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  PREFIX owl: <http://www.w3.org/2002/07/owl#>
3
4  SELECT (str(?label) as ?Termos)
5
6  WHERE
7  {
8    {
9      ?res1 rdfs:label ?label .
10     ?res1 rdfs:subClassOf ?res2 .
11     ?res2 rdfs:label ?o
12     FILTER(lang(?label) = "pt") .
13     FILTER(lcase(?o) = lcase("Doença Bacteriana"@pt)) .
14   }
15   UNION
16   {
17     ?res1 rdfs:label  ?label.
18     ?res1 rdfs:subClassOf ?res2 .
19     ?res2 rdfs:subClassOf ?res3 .
20     ?res3 rdfs:label ?o
21     FILTER(lang(?label) = "pt") .
22     FILTER(lcase(?o) = lcase("Doença Bacteriana"@pt)) .
23   }
24 }
25
```

Fonte: Desenvolvido pelo autor

Ambas as consultas necessitam do prefixo “rdfs”, pois utilizam elementos que definem a estrutura da ontologia como o rótulo (*label*) e a *tag* que indica subclasse (*subClassOf*). Note também que nas linhas 13 e 22 há o termo “Doença Bacteriana”, que foi colocado nestas consultas para exemplificar uma execução. Na aplicação, ao invés do termo há uma variável que recebe um termo extraído da expressão de busca elaborada pelo usuário.

A diferença entre as consultas é que na mais geral busca-se a classe com rótulo igual ao termo digitado e a partir dela obtém-se as classes que a tem como subclasse, enquanto na mais específica, altera-se apenas o ponto de “partida” e a ordem da busca por subclasses.

A cláusula UNION é utilizada para se obter dois níveis de exaustividade, no exemplo em questão, utilizando as classes da ontologia PEDTERM, o primeiro bloco recuperaria o termo “Doença Infecciosa” e o segundo “Doença ou Distúrbio”.

A consulta utilizada para recuperar termos equivalentes pode ser vista na Figura 31. É possível notar que a única diferença dela para as hierárquicas é o uso de uma propriedade diferente, no entanto sua dinâmica de funcionamento é bastante semelhante.

Figura 31 – Consulta SPARQL utilizada para explorar relações de equivalência

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT (str(?label) as ?Termos)

WHERE
{
      ?res1 rdfs:label ?o .
      ?res1 owl:equivalentClass ?res2 .
      ?res2 rdfs:label ?label
      FILTER(lang(?label) = "pt") .
      FILTER(lcase(?o) = lcase("Doença Bacteriana"@pt)) .
}
```

Fonte: Desenvolvido pelo autor

Para que estas consultas possam ser executadas é necessário um processador (ou mecanismo) SPARQL, que é um programa capaz de submete-las a um conjunto de dados local ou remoto e retornar os resultados em algum formato. Quando este um processador SPARQL é instalado e torna-se disponível na rede via requisições HTTP, geralmente identificado por uma URL que permite sua divulgação e acesso, ele passa a ser denominado *Endpoint* SPARQL (IDEHEN, 2018).

Como já discutido o Apache Jena Fuseki implementa uma série de serviços, dentre eles um *Endpoint* SPARQL para cada conjunto de dados em sua base. O ContextOnSearch faz uso deste recurso via requisições HTTP provenientes da linguagem PHP.

Finda-se aqui a apresentação da linguagem SPARQL, pois acredita-se que foram abordados os conceitos concernentes ao entendimento de sua utilização no ContextOnSearch. O SPARQL possui ainda muitas características que não foram abordadas, assim como inúmeras aplicações, porém esta discussão extrapola os objetivos deste estudo.

O próximo tópico aborda a linguagem de programação PHP, que foi a base para o desenvolvimento do ContextOnSearch

5.4.4 PHP e as bibliotecas *composer*, *sparqllib*

A linguagem de programação PHP é voltada ao desenvolvimento de aplicações e sites para a Web, atuando como intermediadora entre os servidores e a interface do usuário. Sua sigla é um acrônimo recursivo para *PHP: Hypertext Preprocessor*.

A linguagem PHP foi criada um propósito bastante simples, no entanto, com o passar dos anos seus propósitos e objetivos amplificados, tornando-a uma linguagem bastante robusta e largamente utilizada (DALL'OGLIO, 2015).

A partir da versão 7 a linguagem passou a suportar uma série de recursos que auxiliaram em sua consolidação como a linguagem mais utilizada em websites, com cerca de 79% de todos os sites (W3TECHS, 2020).

Para sua utilização são necessários três elementos: um interpretador PHP, um servidor Web e um navegador (COWBURN, 1997). Um interpretador é um programa que lê o código fonte de uma linguagem e a executa linha a linha, não gerando um objeto (SILVA, 2017).

No desenvolvimento do ContextOnSearch foi utilizado o Apache HTTP Server[11] com o módulo do interpretador PHP. O Apache HTTP Server é um programa mantido pela *Apache Software Foundation*, uma organização sem fins lucrativos criada para manter projetos de código aberto como o HTTP Server, OpenOffice, Lucene, Hadoop e o Jena Fuseki também utilizado neste projeto (APACHE, 2019).

A sigla HTTP, que faz parte do nome do servidor, é referente ao protocolo *Hypertext Transfer Protocol*, que pode ser traduzido como Protocolo de Transferência de Hipertexto, um protocolo de comunicação responsável pela troca de dados na *World Wide Web* (www) (TANENBAUM, 2003). Ele é importante pois permite que a linguagem PHP se comunique com o ElasticSearch e o Apache Jena Fuseki.

No protótipo são utilizadas ainda duas bibliotecas de *software* para integração do PHP com os serviços do ElasticSearch e do Apache Jena Fuseki. Para comunicação com o Elastic foi utilizado o cliente oficial para a linguagem PHP, fornecido pela própria empresa que

[11] Apache HTTP Server é um servidor HTTP de código aberto utilizado para prover serviços na Internet utilizando o protocolo de comunicação HTTP. Disponível em: https://httpd.apache.org/. Acesso em 19 ago. 2020.

desenvolve a ferramenta. De acordo com a documentação oficial, o cliente PHP para o ElasticSearch é composto por uma série de códigos que facilitam a integração e utilizam métodos de acesso a um API REST (ELASTIC, 2020).

Na comunicação com o Apache Jena Fuseki utilizou-se a biblioteca "sparqllib.php", um componente escrito em linguagem PHP que implementa métodos e rotinas para a execução de consultas ao *SPARQL Endpoint*. Criada no ano de 2012 por Christopher Gutteridge da Universidade de Southampton no Reino Unido, a biblioteca é um *software* livre e assim tem uma licença aberta para utilização sem restrições (GUTTERIDGE, 2012).

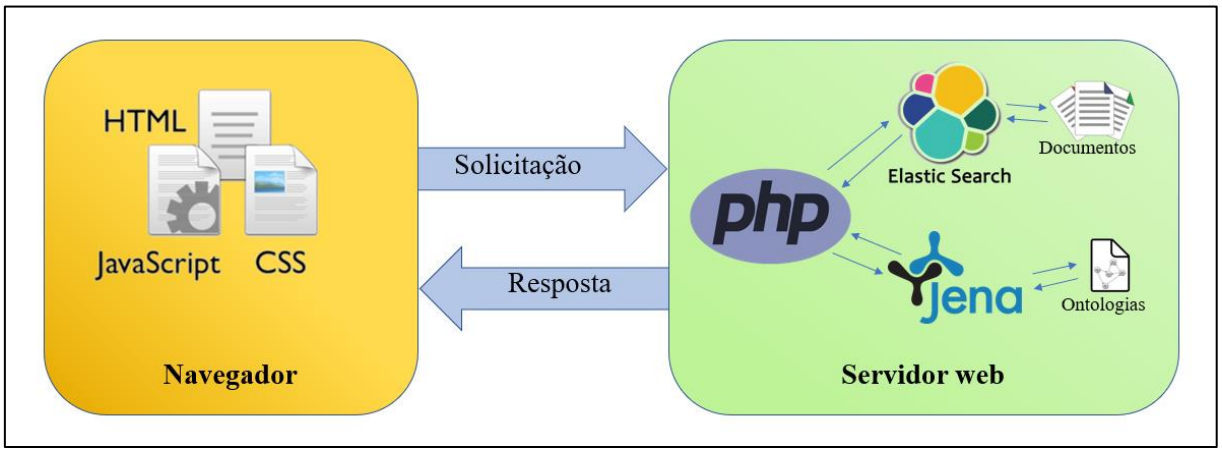
No modelo de arquitetura cliente-servidor os clientes solicitam serviços através do envio de mensagens ao servidor, que por sua vez executa a tarefa devolve uma resposta. O cliente é a parte que interage com o usuário, que fornece uma interface que este desempenhe suas tarefas. Esta é a parte conhecida como *front-end* de uma aplicação, ao passo que o *back-end* é composto pelos processos e serviços que acontecem dentro do servidor para à execução da solicitação do cliente (FILETO, 2006).

No ContextOnSearch a linguagem PHP e os serviços do ElasticSearch e do Apache Jena fuseki compõem o *back end*, enquanto o *front-end* é composto pelas linguagens HTML, CSS, JavaScript e as bibliotecas Bootstrap[12] e jQuery[13]. Uma representação destas tecnologias e como elas se relacionam pode ser vista no diagrama apresentado na Figura 32.

[12] Bootstrap é uma biblioteca *front-end* de código aberto e bastante popular no desenvolvimento de interfaces e componentes visuais em *websites* e outras aplicações. Disponível em: https://getbootstrap.com/. Acesso em 19 ano. 2020.

[13] jQuery é uma biblioteca JavaScript que possibilita a manipulação de elementos, eventos, animações e outras tarefas mais fácil, tornando a programação utilizando a linguagem JavaScript mais simples e ágil. Disponível em: https://jquery.com/. Acesso em 19 ago. 2020.

Figura 32 – Diagrama das tecnologias utilizadas no ContextOnSearch



Fonte: Desenvolvido pelo autor

5.4.5 Linguagens HTML, CSS e JavaScript

Dentre as linguagens para criação do *front-end* de *websites* e aplicações Web as mais conhecidas são HTML, CSS e JavaScript. Elas formam uma pilha de tecnologias complementares, pois, implementam respectivamente estrutura, apresentação e comportamento de um *software* Web.

A linguagem HTML não é uma linguagem de programação, ela é classificada como uma linguagem de marcação, pois sua função é estruturar e organizar os conteúdos de forma padronizada para que os programas que interpretam essas páginas sejam capazes de apresentar adequadamente as informações em uma página Web.

O acrônimo HTML vem do inglês *HyperText Markup Language*, que em português pode ser traduzido como Linguagem de Marcação de Hipertexto. Criada por Tim Berners-Lee no final dos anos oitenta como um projeto de troca de informações utilizando a *World Wide Web* (www), um sistema de troca de informações interligado e executado pela Internet (W3C, 1992).

Outro elemento que compõem a interface da aplicação é a CSS, uma linguagem utilizada para definir a aparência de documentos construídos com linguagens de marcação como o HTML. CSS é o acrônimo para *Cascading Style Sheets* ou Folhas de Estilo em Cascata, em português. Sua função é definir como serão exibidos para o usuário os elementos e seus respectivos conteúdos (BOS, 2016).

A primeira recomendação de uso da CSS pelo W3C aconteceu em 1996, motivada pela necessidade de estilização das inúmeras páginas Web que estavam surgindo na época (BOS,

2016). Juntos, HTML e CSS possibilitam a criação de páginas Web completas, que são capazes de apresentar informações nos mais diversos formatos e layouts.

O último elemento é o JavaScript (JS), que teve sua origem na necessidade de implementação de ações e interações mais dinâmicas nas páginas Web. Páginas HTML estilizadas com CSS são por sua natureza estáticas, o JS possibilitou o desenvolvimento de diversos comportamentos web websites conhecidos atualmente.

Funções como menus deslizantes, *zoom* em imagens, ou simplesmente o ajuste de uma página conforme a orientação do dispositivo e/ou o tamanho da tela são exemplos comportamentos úteis, e que proporcionam uma boa usabilidade, que podem ser implementados com JavaScript. A adição desta camada de comportamentos possibilitou mudanças profundas no desenvolvimento Web ao longo dos anos.

É importante observar que a linguagem JavaScript nada tem a ver com linguagem Java, embora compartilhem parte do nome e possam ser utilizados no paradigma orientado a objetos elas diferem em quase todo o restante, inclusive em relação à função de cada uma.

A combinação destas três tecnologias é suficiente para a criação de uma interface funcional e usável para uma aplicação Web, porém, com o propósito de facilitar o desenvolvimento foram utilizadas duas bibliotecas no ContextOnSearch: *Bootstrap* e *jQuery*.

Boostrap é uma biblioteca de componentes de interface pré-construídos para o desenvolvimento de aplicações Web com HTML, CSS e JS. Possui código fonte aberto e é considerada a biblioteca mais popular do mundo em seus propósitos (BOOTSTRAP, 2020).

Já o *jQuery* é uma biblioteca específica de JS que implementa recursos como manipulação de documentos, eventos, animações e controle de requisições assíncronas em páginas Web (JQUERY, 2021). Seus principais benefícios são a facilidade em programar e a padronização dos métodos entre diferentes navegadores.

O produto final da utilização de todas as tecnologias apresentadas é mostrado na próxima seção, que aborda os resultados obtidos a partir de uma série de consultas submetidas ao sistema.

5.5 Apresentação e Análise dos resultados obtidos

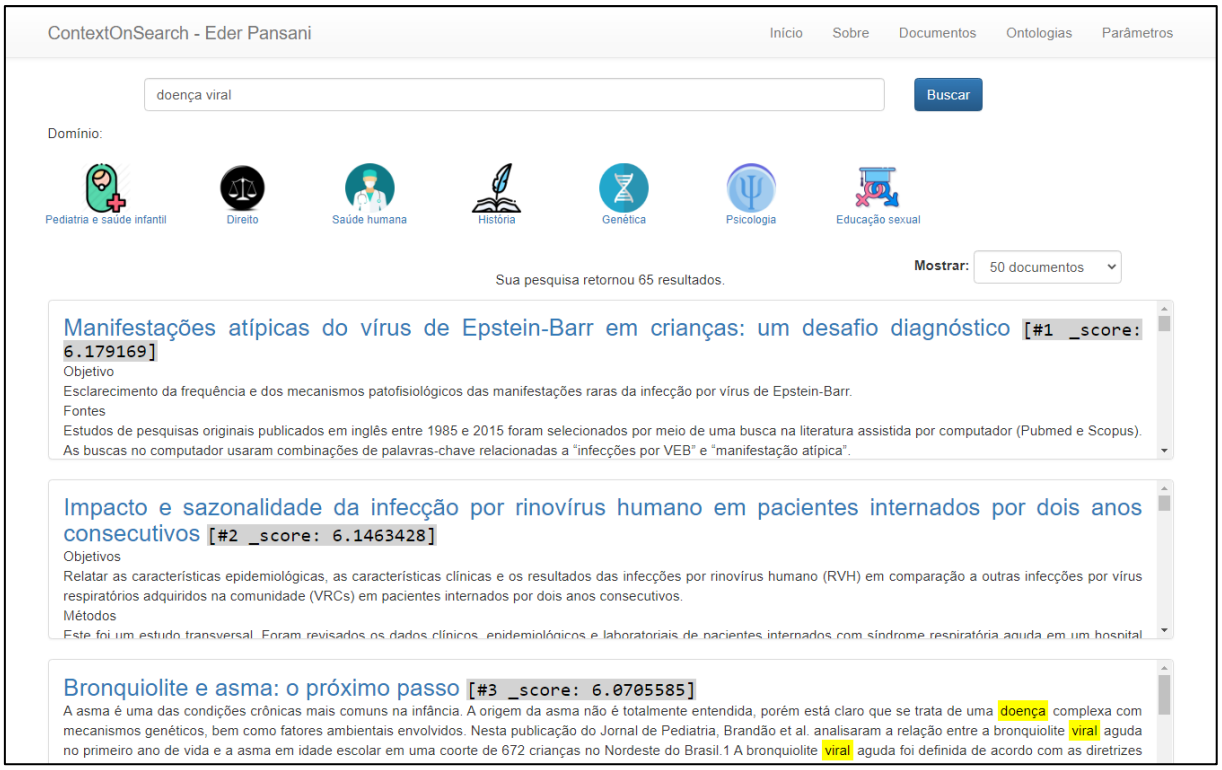
O ContextOnSearch é um mecanismo de busca baseado em uma caixa de texto livre. A apresentação dos resultados das buscas é feita uma lista de documentos ordenada pela relevância calculada entre a consulta e os documentos da coleção.

Para facilitar a apresentação e a discussão dos resultados obtidos optou-se por utilizar o recurso de realce (*highlight*), que permite destacar palavras ou trechos de texto de cada um dos registros recuperados em uma busca em relação à consulta (GORMLEY; TONG, 2015, p.19, tradução nossa).

Outra decisão que também teve como objetivo facilitar a compreensão dos resultados foi a apresentação do valor de relevância (*score*) de cada um dos documentos recuperados. No ElasticSearch a pontuação de relevância é um número de ponto flutuante que determinada a classificação do item no conjunto de resultados (GORMLEY; TONG, 2015, p. 111, tradução nossa).

A Figura 33 apresenta a página de busca do ContextOnSearch com os recursos de *highlight e score* habilitados. O valor de relevância (*score*) é mostrado após o título do registro enquanto o realce é destacado em amarelo em qualquer parte do documento, incluindo o título.

Figura 33 – Página de busca do ContextOnSearch com destaque aos valores de realce e relevância



Fonte: Desenvolvido pelo autor

Ainda na Figura 33 pode-se ver, na parte superior, a caixa de texto e o botão “Buscar” à sua direita. Utilizando estes campos o usuário elabora sua expressão de busca e a submete ao sistema. A quantidade de resultados recuperados seguida de cada um dos documentos é apresentada logo após a lista de domínios encontrados.

Os domínios que podem ser vistos em destaque na Figura 34 representam as ontologias existentes no ContextOnSearch e que possuem um ou mais termos da expressão de busca. Ressalta-se que com exceção do item “Pediatria e saúde infantil”, oriundo da ontologia PEDTERM, os demais são fictícios e não representam ontologias reais, eles foram inseridos no sistema apenas com o propósito de demonstrar o comportamento do *software* com diversos domínios.

Figura 34 – Destaque à área de contextualização da expressão de busca

Fonte: Desenvolvido pelo autor

A partir desta lista de domínios, o usuário pode executar a segunda etapa do método, a **contextualização da busca**, que consiste na escolha da opção que melhor representa o contexto de suas necessidades informacionais. Esta etapa é facultativa, sendo que caso o utilizador sinta-se satisfeito com o conjunto de resultados recuperados ele pode acessá-los e finalizar a utilização do buscador.

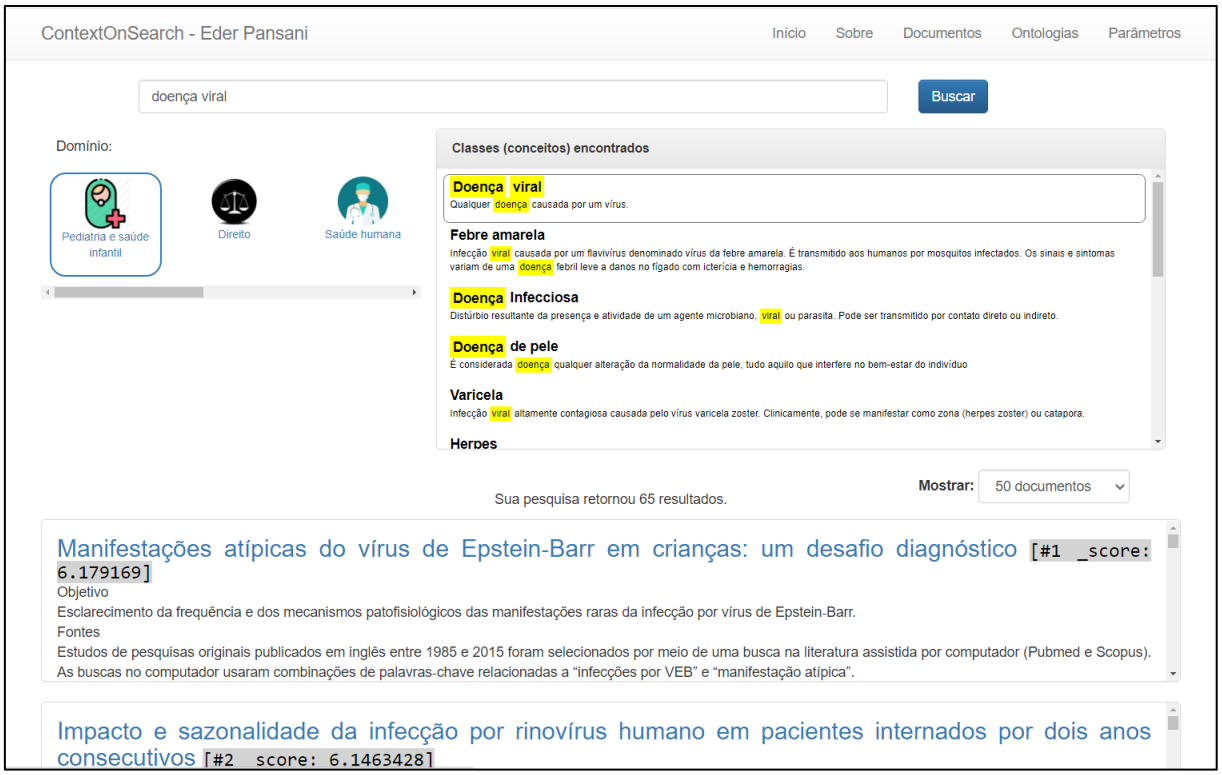
Quando o usuário faz a escolha de um dos domínios da lista, ele está inconscientemente escolhendo a ontologia que fornecerá subsídios à etapa de expansão da consulta por meio da identificação de termos relacionados.

No entanto, em uma mesma ontologia podem existir diferentes conceitos que correspondam total ou parcialmente aos termos da expressão de busca, sendo assim, é necessária uma etapa de **determinação do conceito** que mais se aproxima do conceito central da expressão de busca.

Na prática, a escolha de um dos conceitos apresentados na interface do mecanismo informa ao sistema qual será a classe inicial do processo de identificação de termos relacionados, que possibilita a construção do diagrama hierárquico com os conceitos relacionados.

A Figura 35 destaca a área de apresentação dos conceitos que correspondem total ou parcialmente aos termos da expressão de busca. É possível observar que a listagem dos conceitos está ordenada por sua proximidade com os termos da consulta.

Figura 35 – Destaque à área de escolha da classe (conceito) inicial



Fonte: Desenvolvido pelo autor

A etapa seguinte é a **identificação e apresentação dos conceitos relacionados** à expressão de busca com suas relações em um diagrama hierárquico, como mostrado na Figura 36. O propósito da apresentação deste diagrama é proporcionar ao usuário um recurso que pode auxiliar em sua familiarização com o domínio do conhecimento pesquisado, o contexto de sua busca.

Por meio desta hierarquia o usuário pode ainda ter acesso à terminologia do domínio, tendo assim condições de escolher termos úteis ao aprimoramento de sua consulta e consequentemente ter acesso a resultados mais relevantes.

Figura 36 – Destaque à área de apresentação da hierarquia de termos relacionados

Fonte: Desenvolvido pelo autor

O último passo do método de **expansão da consulta** é a escolha dos termos que serão adicionados e/ou removidos. Esta ação é completamente manual, ficando a cargo do usuário a escolha dos termos que lhe pareçam mais adequados. No diagrama hierárquico o termo escolhido inicialmente é apresentado em destaque, para que o usuário tenha condições de observar outros termos genéricos ou específicos em relação a ele.

Para adicionar um termo o usuário deve clicar/tocar sobre ele, o sistema irá adicioná-lo à caixa de pesquisa colocando-o entre aspas para diferenciá-lo dos termos anteriormente descritos. Quando desejar o usuário pode refazer a busca utilizando a nova expressão, iniciando assim uma nova iteração no buscador.

No exemplo apresentado na Figura 37 o usuário escolheu o termo relacionado "Gripe", que foi adicionado à consulta e resultou na recuperação de 69 documentos. A busca inicial, que pode ser observada na Figura 33 retornava 65 registros, indicando assim um aumento na revocação do sistema.

Analisando a Figura 37 pode-se notar ainda que houveram alterações na classificação dos resultados, pois, a partir da consulta expandida o ContextOnSearch passa a considerar mais

relevantes os documentos que contém os termos “doença viral” e “Gripe” alterando assim a ordenação dos resultados de busca.

Figura 37 – Exemplo de uma busca com a consulta “doença viral” expandida

Fonte: Desenvolvido pelo autor

Essa diferença na ordenação de relevância pode ser observada comparando os valores de “*score*” dos primeiros resultados de busca da Figura 33 e Figura 37, respectivamente 6.179169 e 12.776826. Essa diferença se dá em razão da quantidade de ocorrências dos termos no corpo e título dos documentos, que influência no cálculo do *score*.

Considere agora uma situação hipotética em que o usuário que buscava informações sobre as doenças virais deseja saber mais sobre a doença conhecida como “catapora”. Ele segue os mesmos passos, escolhendo agora o termo “catapora” para expandir sua consulta.

O resultado desta busca é mostrado na Figura 38, onde é possível notar que a quantidade de documentos recuperados permanece a mesma, no entanto, a classificação de relevância sofreu alterações. Isso significa que o termo “catapora” não o descritor mais adequado para recuperar informações sobre essa doença.

Figura 38 – Exemplo de busca expandida com o termo “catapora”

Fonte: Desenvolvido pelo autor

Observando a hierarquia ter termos relacionados pode-se notar que “catapora” e “varicela” possuem entre si uma linha horizontal tracejada, ligando-os e indicando ao usuário que os termos são equivalentes.

A partir desta percepção o usuário pode então adicionar o termo “varicela” à sua consulta, situação que resulta em um conjunto de 68 registros recuperados, como pode-se ver na Figura 39. Assim como no exemplo anterior a adição de um termo à expressão de busca influência no cálculo da relevância (valor de *score*), o que não provoca alterações apenas nos dois primeiros registros recuperados.

O recurso de destaque aos termos distintos que possuem o mesmo significado, ou seja, remetem ao mesmo conceito pode promover benefícios significativos ao processo de recuperação de informação, pois auxilia o usuário a superar alguns dos problemas linguístico-terminológicos inerentes às linguagens.

Outro benefício do método proposto reside na apresentação do diagrama hierárquico com os termos relacionados, que é uma estratégia de apresentação da terminologia do domínio representada na ontologia de forma visual e acessível.

Figura 39 – Exemplo de busca expandida com o termo “catapora” e “varicela”

Fonte: Desenvolvido pelo autor

A dinâmica simples de adição de termos para sucessivas buscas também pode ser um fator decisivo, pois permite ao usuário sucessivas interações com o mecanismo de busca sem demandar um investimento de tempo ou esforço significativos.

O diagrama hierárquico de termos permite, em razão de sua estrutura em níveis, que os usuários sejam capazes de identificar os termos superiores e inferiores ao conceito inicial, possibilitando a especificação ou generalização das buscas. Tais benefícios podem conduzir a um conjunto de resultados supostamente menor e mais preciso ou maior e mais abrangente respectivamente. Ficando a critério do usuário o tipo de expansão que lhe pareça mais capaz de satisfazer suas necessidades informacionais.

Um último recurso que pode ser observado na Figura 39 logo abaixo da caixa de busca é uma mensagem seguida de um *hiperlink* para a consulta inicial. Este controle possibilita ao usuário restaurar o mecanismo de busca à consulta inicialmente formulada e seus respectivos resultados, desconsiderando a contextualização e desfazendo as expansões.

Os métodos de contextualização e expansão de consultas do ContextOnSearch foram implementados considerando questões de usabilidade, demandando poucas interações com o sistema e utilizando elementos visuais e gráficos para promover maior facilidade de utilização.

No entanto, é importante salientar que não são exploradas as recomendações da área do design de interfaces ou da experiência de usuário (*user experience*), limitando o conceito de usabilidade a capacidade de um usuário que já utilizou algum mecanismo de busca Web ser capaz de utilizar o sistema sem um treinamento prévio.

Com base nas percepções pessoais obtidas com a utilização do protótipo desenvolvido, visto que não foram feitos testes com usuários, foi possível notar que a qualidade da expressão de busca em um sistema de recuperação de informação tem relação direta com a revocação e a relevância dos resultados apresentados.

O ContextOnSearch atende às propostas de melhorias em SRIs indicadas por Jackson e Moulinier (2007), Woods (2004), Hyvönen *et al*. (2005), Lei, Uren e Motta (2006), Beppler (2008), Baeza-Yates e Ribeiro-Neto (2013) e Shneiderman *et al*. (2017) disponibilizando em formato visual uma hierarquia de termos do domínio de interesse do usuário, obtidos após a contextualização da necessidade de informação do usuário, apresentando a terminologia específica presente na ontologia.

Disponibiliza-se ao usuário, em um formato visual, os conceitos e suas relações para que ele tenha melhores condições de descrever sua consulta, possibilitando assim a elaboração de expressões de busca mais significativas e menos ambíguas, que por sua vez podem conduzir a resultados de busca mais relevantes.

6 Considerações finais

O objetivo geral deste trabalho foi a proposição de um método de contextualização das necessidades informacionais e da expansão de consultas utilizando termos identificados a partir das relações semânticas de uma ontologia de domínio, em sistemas de recuperação de informação no ambiente Web. Sua proposição teve origem na análise dos principais problemas que interferem na tarefa de encontrar informação utilizando mecanismos de busca, que indicou desafios relacionados à elaboração da expressão de busca.

A investigação partiu do pressuposto que a contextualização, utilizando a estrutura terminológica de uma ontologia, e a expansão da consulta utilizando as relações semânticas entre os conceitos das ontologias proporcionam melhores condições de representar as necessidades informacionais dos usuários de um mecanismo de busca, proporcionando maior eficiência do processo de recuperação de informação. Levantou-se ainda a hipótese de que a apresentação da hierarquia de conceitos relacionados à expressão de busca favorece ao usuário a compreensão da estrutura terminológica do domínio de interesse.

Com base neste cenário, as primeiras etapas do estudo consistiram de uma revisão de literatura sobre os principais problemas envolvidos no processo de busca e recuperação de informação, relacionados com os aspectos de comunicação entre o usuário e o sistema. Neste sentido, para que o mecanismo de busca “compreenda” o domínio ou o contexto em que se insere a necessidade de informação do usuário é necessário que este interaja ativamente com o sistema por meio de uma interface adequada.

Durante o desenvolvimento desta tese foram analisados trabalhos com abordagens que consistem na utilização de ontologias para o aprimoramento da recuperação de informação. Entretanto, não foram encontradas referências a abordagens que as utilizam para a criação de recursos visuais e interativos.

Nesta pesquisa foi utilizada a estrutura terminológica das ontologias e suas relações explícitas de hierarquia e equivalência entre conceitos. No entanto, as ontologias possuem um potencial de uso muito maior, como a possibilidade de utilização suas restrições e atributos na identificação de conceitos ou a geração de novas informações por meio da inferência.

O resultado mais significativo da pesquisa foi a proposição do “método de contextualização e expansão de consultas baseado em ontologias de domínio”, idealizado e

desenvolvido a partir das perspectivas da recuperação a informação e do uso das ontologias na organização da informação. Apesar de sua idealização ter sido focada em mecanismos de busca no ambiente Web, acredita-se que ele tem potencial de ser aplicado em múltiplos contextos, tais como: mecanismos de busca multimídia, repositórios institucionais, entre outros.

O diferencial do método proposto está na interação do usuário com o mecanismo de busca, permitindo uma comunicação mais efetiva, e com isso, possibilitando a elaboração de expressões de busca que representam as necessidades de informação de forma mais precisa, e por consequência, obtendo resultados mais relevantes.

Outro resultado foi a implementação de um protótipo de *software*, denominado ContextOnSearch, que consiste em um mecanismo de busca que implementa o método proposto. Sua função é demonstrar a utilização do método em um mecanismo de busca real, servindo como uma prova de conceito.

Para sua implementação foram empregadas diversas tecnologias, dentre elas as linguagens de programação e marcação para Web: HTML, CSS, JavaScript e PHP. O Apache Jena Fuseki foi utilizado como motor de inferências e a linguagem SPARQL como protocolo de consulta. A base do buscador foi o ElasticSearch, um *software* de código aberto construído sobre o Apache Lucene. Com isso foi possível concluir que o uso de ferramentas livres pode resultar em soluções com amplo potencial de aprimoramento do processo de busca e recuperação de informação.

Para a execução de testes e análise do funcionamento do método no ContextOnSearch foi utilizada a ontologia *Pediatric Terminology* (PEDTERM), que trata da pediatria e saúde infantil. Foi criada ainda uma coleção de documentos oriundos de um periódico nacional sobre a mesma temática, caracterizando um ambiente controlado de um único assunto.

A partir da análise qualitativa, realizada com os resultados de uma série de consultas submetidas ao ContextOnSearch, foi possível perceber melhorias, como o aumento da revocação sem perdas significativas na precisão. Além de melhorias na classificação (ordenação por relevância) dos resultados, posicionando documentos com informações mais úteis dentre as primeiras posições no *ranking*.

Embora muitos resultados tenham sido positivos, foram identificados também problemas que podem limitar o uso do método. Dentre eles estão a dependência de ontologias dotadas de rótulos em suas classes e a dificuldade em lidar com termos complexos, isto é,

expressões de busca formadas por duas ou mais palavras. Nesse sentido foram observados resultados positivos em consultas com até 3 palavras e negativos em consultas com mais de 5 palavras.

A dificuldade em lidar com consultas longas é uma falha específica do ContextOnSearch, ocasionada por sua limitação em buscar na ontologia por múltiplos termos/conceitos relacionados a uma única expressão de busca. No entanto, ressalta-se que o método proposto é capaz de lidar com múltiplos conceitos dentro de uma expressão de busca. Desta forma, o aprimoramento do ContextOnSearch para superar essa limitação é uma oportunidade de pesquisa a ser explorada.

Durante o desenvolvimento do protótipo não foram observados os conceitos e recomendações do Design de Interfaces, da Interação Humano computador e da Arquitetura da Informação, imprescindíveis à implementação de uma solução de *software* consistente na atualidade. Este fato apresenta-se como uma oportunidade de pesquisas futuras, que podem ainda englobar análises e avaliação de aspectos de desempenho e performance, como tempo de resposta em grandes acervos.

Outra possibilidade de pesquisas futuras está na execução de testes com usuários, buscando avaliar se os recursos de contextualização e expansão de consultas melhoram efetivamente a recuperação de informação relevante.

Apesar das limitações, comprova-se a tese de que o uso de um método de contextualização e expansão de termos de busca utilizando ontologias de domínio e seus relacionamentos semânticos auxilia na tarefa de formulação da consulta, contribuindo com a melhoria no processo de recuperação de documentos textuais.

Pretende-se publicar os resultados desta pesquisa em periódicos e eventos acadêmicos para possibilitar a socialização e discussão de suas contribuições, que em âmbito científico podem contribuir com o desenvolvimento da área de Recuperação da Informação. No contexto técnico e profissional acredita-se que os resultados podem ser úteis ao desenvolvimento de mecanismos e sistemas de busca mais adequados à tarefa de elaboração da expressão de busca.

Por fim, em âmbito social, os resultados podem contribuir com a modificação e evolução na dinâmica de recuperar informação, por meio da disponibilização de recursos interativos nos sistemas de recuperação de informação que auxiliam as pessoas a traduzir suas necessidades de

informação em expressões de consulta mais eficientes e significativas, que conduzam a respostas mais adequadas e relevantes aos seus questionamentos.

Acredita-se que a contextualização das necessidades informacionais dos usuários seja a maior contribuição deste estudo, que possibilita a exploração de abordagens como a expansão das consultas utilizando termos relacionados e outras a serem estudadas por pesquisas vindouras. O aprimoramento do processo de representação das necessidades informacionais dos usuários é imprescindível para uma recuperação mais eficiente e eficaz, pois fortalece o "elo mais frágil" do processo de RI, o usuário, com suas dificuldades e limitações.

REFERÊNCIAS

AGUADO, G. *et al*. ONTOGENERATION: Reusing domain and linguistic ontologies for Spanish text generation. *In*: WORKSHOP ON APPLICATIONS OF ONTOLOGIES AND PROBLEM SOLVING METHODS, 13., 1998, Brighton – UK. **Anais** […]. Brighton: ECAI, 1998. Disponível em: http://oa.upm.es/6449/1/ONTOGENERATION__Reusing_.pdf. Acesso em: 11 dez. 2019.

ALLAN, J.; CARTERETTE, B.; LEWIS, J. When will Information Retrieval be "Good Enough?". *In*: PROCEEDINGS OF THE 28TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 28., 2005, New York. **Anais** […]. New York: ACM. 2005. p. 433-440, Disponível em: https://dl.acm.org/doi/abs/10.1145/1076034.1076109. Acesso em: 12 abr. 2021.

ALMEIDA, D. P. R. *et al*. Paradigmas contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. **Revista Eletrônica Informação e Cognição**, Marília, v. 6, n. 1, p. 16-27, 2007. DOI: https://doi.org/10.36311/1807-8281.2007.v6n1.745 Disponível em: https://revistas.marilia.unesp.br/index.php/reic/article/view/745. Acesso em: 02 abr. 2021.

AMATO, G. *et al*. Large-scale image retrieval with elasticsearch. *In*: ACM SIGIR CONFERENCE ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR), 41., 2018, Ann Arbor – Michigan. **Anais** […]. New York: Association for Computing Machinery, 2018. p. 925-928. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3209978.3210089. Acesso em: 28 jun. 2021.

ANDRADE, M. T. T.; FERREIRA, C. V.; PEREIRA, H. B. B. Uma ontologia para a Gestão do Conhecimento no Processo de Desenvolvimento de Produto. **Gest. Prod.**, São Carlos, v. 17, n. 3, p. 537-551, 2010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2010000300008&lng=en&nrm=iso. Acesso em: 23 jan. 2020. DOI: http://dx.doi.org/10.1590/S0104-530X2010000300008.

ANDREOU, A. **Ontologies and query expansion**. 2005. 81f. Dissertação (Mestrado) - Universidade de Edimburgo, Edimburgo, 2005. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.2802&rep=rep1&type=pdf. Acesso em: 28 jun. 2021.

ANTONIOU, G.; FRANCONI, E.; HARMELEN, F. V. Introduction to semantic web ontology languages. *In*: **Reasoning web**. Springer, Berlin, Heidelberg, 2005. p. 1-21. Disponível em: https://link.springer.com/chapter/10.1007/11526988_1. Acesso em: 05 dez. 2019.

APACHE Jena Fuseki. [201-]. Documentação. Disponível em: https://jena.apache.org/documentation/fuseki2/. Acesso em: 27 jul. 2020.

APACHE, WHAT IS THE ASF? **apache.org**, 2019. Disponível em: http://www.apache.org/foundation/. Acesso em: 17 jul. 2020.

APPOLINÁRIO, F. **Metodologia da ciência**: filosofia e prática da pesquisa. 2. ed. São Paulo: Cengage Learning, 2011.

ARAUJO, V. M. A. P. Sistemas de recuperação da informação: uma discussão a partir de parâmetros enunciativos. **Transinformação**, v. 24, n. 2, p. 137-143, 2012. DOI: http://dx.doi.org/10.1590/S0103-37862012000200006. Disponível em: https://www.scielo.br/scielo.php?pid=S0103-37862012000200006&script=sci_abstract&tlng=es. Acesso em: 23 mar. 2021.

AULETE DIGITAL. **Dicionário Aulete Digital**. [2014]. Disponível em: https://www.aulete.com.br/. Acesso em: 28 out. 2019.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. 2.ed. Porto Alegre: Bookman, 2013. 590 p.

BAPTISTA, F. **Uma proposta de interface de resultados de busca em sistemas de recuperação de informação**: a Semiótica e a Interação Humano Computador como aporte teórico. 2019. 173 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, Marília, 2019.

BARONAS, J. E. A. Variação linguística na escola: propostas de ação. **SIGNUM: Estud. Ling.**, Londrina, n. 14/2, p. 105-116, dez. 2011.

BARQUIN, B. A. R.; GONZÁLEZ, J. A. M.; PINTO, A. L. Construção de uma ontologia para sistemas de informação empresarial para a área de telecomunicações. **DataGramaZero/Rev. Ci. Inf.**, Brasília, v. 7, n. 2, 2006. Disponível em: http://www.brapci.inf.br/_repositorio/2010/01/pdf_f5f8bd5054_0007557.pdf. Acesso em: 03 mar. 2021.

BARROS, J. A. **Os conceitos**: seus usos nas ciências humanas. Petrópolis - RJ: Editora Vozes, 2016.

BASILIO, M. Em torno da palavra como unidade lexical: palavras e composições. **Veredas**: Revista de Estudos Linguísticos, Juiz de Fora, v. 4, n. 2. p. 9-18, jul. 2016. Disponível em: https://periodicos.ufjf.br/index.php/veredas/article/view/25314. Acesso em: 28 jun. 2021.

BAX, M. P.; COELHO, E. M. P. Compromissos ontológicos e pragmáticos em ontologias informacionais: convergências e divergências. **DataGramaZero - Revista de Informação**, v. 13, n. 3, p. 1-10. 2012. Disponível em: https://brapci.inf.br/index.php/article/download/53311. Acesso em: 24 abr. 2021.

BECHHOFER, S. *et al*. **OWL web ontology language reference**. World Wide Web Consortium (W3C), 2004. Disponível em: https://www.w3.org/TR/owl-ref/. Acesso em: 28 jun. 2021.

BEPPLER, F. D. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. 2008. 123 f. Tese (Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2008.

BHAGDEV, R. *et al*. Hybrid search: Effectively combining keywords and semantic searches. *In*: **European semantic web conference**. Springer, Berlin, Heidelberg, 2008. p. 554-568.

BHOGAL, J.; MACFARLANE, A.; SMITH, P. A review of ontology-based query expansion. **Information Processing and Management**, v. 43, n. 4, p. 866-886, 2007.

BISCALCHIN, R.; BOCCATO, V. R. C. Estudo do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias: uma abordagem qualitativa-sociocognitiva pela perspectiva do usuário. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 24., 2011, Maceió. **Anais** [...]. Maceió: CBBD, 2011. Disponível em: http://febab.org.br/congressos/index.php/cbbd/xxiv/paper/view/39/508. Acesso em: 03 mar. 2021.

BONINO, D. *et al*. Ontology driven semantic search. **WSEAS Transaction on Information Science and Application**, v. 1, n. 6, p. 1597-1605, 2004.

BOOTSTRAP. **Build fast, responsive sites with Bootstrap**, [2020?]. Disponível em: https://getbootstrap.com. Acesso em: 20 mar. 2021.

BOS, B. **A brief history of CSS until 2016**. World Wide Web Consortium (W3C), 2016. Disponível em: https://www.w3.org/Style/CSS20/history.html. Acesso em: 20 mar. 2021.

BOURAMOUL, A. Contextualisation of information retrieval process and document ranking task in web search tools. **International Journal of Space-Based and Situated Computing**, United Kingdom, v. 6, n. 2, p. 74-89, 2016. DOI: https://doi.org/10.1504/IJSSC.2016.077970. Disponível em: https://www.inderscienceonline.com/doi/abs/10.1504/IJSSC.2016.077970. Acesso em: 23 mar. 2021.

BRASCHER, M.; CAFÉ, L. Organização da informação ou organização do conhecimento? *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 9., 2008, São Paulo. **Anais** [...]. São Paulo: ECA/USP, ENANCIB, 2008. Disponível em: http://enancib.ibict.br/index.php/enancib/ixenancib/paper/view/3016. Acesso em: 21 dez. 2018.

BRÄSCHER, M.; CARLAN, E. Sistemas de organização do conhecimento: antigas e novas linguagens. *In*: ROBREDO, J.; BRÄSCHER, M. (org.). **Passeios no bosque da informação**: estudos sobre representação e organização da informação e do conhecimento – EROIC. Brasília-DF: IBICT, 2010, p. 147-176 Disponível em: https://www2.senado.leg.br/bdsf/handle/id/189812. Acesso em: 23 mar. 2021.

BRICKLEY, D; GUHA, R. V., **RDF Schema 1.1**, World Wide Web Consortium (W3C), 2014. Disponível em: https://www.w3.org/TR/rdf-schema/. Acesso em: 18 mar. 2021.

BROUGHTON, V. *et al*. Knowledge organization. *In*: KAJBERG, L.; LORRING, L. (ed.). **European Curriculum Reflections on Library and Information Science Education**. Copenhagen: Royal School of Library and Information Science, 2004. p. 133-148. Disponível em: http://hdl.handle.net/10150/105851. Acesso em: 23 mar. 2021.

BUCKLAND, M. K. Information as thing, **Journal of the American Society of Information Science**, v.42, n.5, 1991. p.351-360.

BULCAO-NETO, R. F.; PRAZERES, C. V. S.; PIMENTEL, M. G. C. Web Semântica: Teoria e Prática. *In*: TEIXEIRA, C. A. C. *et al*. (org.). **Tópicos em Sistemas Interativos e Colaborativos**. Natal: SBC, 2006. p. 1-40.

BUSCH, J. E. *et al*. **Ontology-based parser for natural language processing**. U.S. Patent n. 7,027,974, 11 abr. 2006. Disponível em: https://patents.google.com/patent/US7027974B1/en. Acesso em: 19 dez. 2019.

BÜTTCHER, S.; CLARKE, C. L. A.; CORMACK, G. V. **Information retrieval**: Implementing and evaluating search engines. Mit Press, 2010.

CABRÉ, M. T. **La terminologia**: teoría, metodología, aplicaciones. Barcelona: Editorial Antártida/Empúries, 1993.

CAMACHO, R. G. Norma culta e variedades lingüísticas. *In*: CECCANTINI, J. L. C. T.; PEREIRA, R. F. (Eds.). **Cadernos de formação**: língua portuguesa. São Paulo: Prograd UNESP, 2004. p. 47-60.

CAMPOS, M. L. A. Ontologias e definições: a explicitação do compromisso ontológico. **ISKO Brasil**, v. 2, p. 132-140. Disponível em: https://brapci.inf.br/index.php/res/v/135103. Acesso em: 24 abr. 2021.

CAMPOS, M. L. A.; CAMPOS, L. M.; MEDEIROS, J. S. A representação de domínios de conhecimento e uma teoria de representação: a ontologia de fundamentação. **Informação & Informação** (UEL. Online), v. 16, n. 2, p. 140-164, 2011.

CAPRA, R. G.; PÉREZ-QUIÑONES, M. A. Using web search engines to find and refind information. **IEEE Computer Society**, v.38, n.10, p.36-42. 2005.

CAPURRO, R. Epistemologia e ciência da informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 5., 2003. Belo Horizonte. **Anais** [...]. Belo Horizonte: UFMG, 2003.

CARDENAS, Y. G. **Modelo de ontologia para representação de jogos digitais de disseminação do conhecimento**. 2014. 149f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2014. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/129486. Acesso em: 03 mar. 2021.

CARLAN, E. **Ontologia e Web semântica**. 2006. 61f. Monografia (graduação) – Universidade de Brasília, Curso de Graduação em Biblioteconomia, 2006. Disponível em: http://eprints.rclis.org/15176/1/ECarlan.pdf. Acesso em: 03 mar. 2021.

CARLAN, E. **Sistemas de organização do conhecimento**: uma reflexão no contexto da ciência da informação. 2010. 195f. Dissertação (mestrado) – Universidade de Brasília, Departamento de Ciência de Informação e Documentação, Brasília, 2010. Disponível em: https://repositorio.unb.br/handle/10482/7465. Acesso em: 03 mar. 2021.

CASTELLS, P.; FERNANDEZ, M.; VALLET, D. An adaptation of the vector-space model for ontology-based information retrieval. **IEEE transactions on knowledge and data engineering**, v. 19, n. 2, p. 261-272, 2006. Disponível em: https://ieeexplore.ieee.org/abstract/document/4039288/. Acesso em: 03 mar. 2021.

CASTRO, S. **Ontologia**. Rio de Janeiro: Jorge Zahar, 2008.

CAVALCANTI, E. P. Revolução da informação: algumas reflexões. **Cadernos de Pesquisas em Administração-Programa de Pós-Graduação em Administração da FEA/USP**, v. 1, n. 01, p. 40-46, 1995.

CERVANTES, B. M. N. Identificação e representação de conceitos por meio de termos. **Informação & Sociedade: Estudos**, v. 30, n. 4, p. 1-14, 2020. DOI: 10.22478/ufpb.1809-4783.2020v30n4.57261. Disponível em: https://brapci.inf.br/index.php/res/v/153386. Acesso em: 23 mar. 2021.

CESARINO, M. A. N. Sistemas de recuperação da informação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 157-168, 1985. Disponível em: https://brapci.inf.br/index.php/article/view/0000009051. Acesso em: 24 abr. 2021.

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

CLEVERDON, C. W. The Cranfield tests on index language devices. **ASLIB Proceedings**. London. v. 19, n. 6, p. 173-194, jun. 1967. DOI: https://doi.org/10.1108/eb050097. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/eb050097/full/html. Acesso em: 23 mar. 2021.

CONEGLIAN, C. S. **Modelo computacional de recuperação da informação para repositórios digitais utilizando ontologias**. 2017. 145f. Dissertação (Mestrado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017.

COWBURN, P. **O que o PHP pode fazer?**, PHP Documentation Group, c1997-2021. Disponível em: https://www.php.net/manual/pt_BR/intro-whatcando.php. Acesso em: 19 mar. 2021.

CUNHA, I. M. R. F. Análise documentária. *In*: SMIT, J. W. (Coord.) - **Análise documentária**: a análise da síntese. 2. ed. Brasília: IBICT, 1989. p. 39-62.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 2, p. 101-107, 1978. Disponível em: http://revista.ibict.br/ciinf/article/view/115. Acesso em: 02 mar. 2021.

DALL'OGLIO, P. **PHP Programando com Orientação a Objetos**. 3. ed. São Paulo: Novatec, 2015.

DANTAS, C. S.; SILVA, T. V. G.; SOUZA, A. C. B. Processo de recuperação da informação: barreiras encontradas pelos usuários. **Biblionline**, v. 9, n. 1, p. 16–26, 2013.

DEL-MASSO, M. C. S.; COTTA, M. A. C.; SANTOS, M. A. P. **Ética em Pesquisa Científica**: conceitos e finalidades. Redefor Educação Especial e Inclusiva - Texto II. São Paulo: Unesp, p. 1-16, 2014. Disponível em: http://acervodigital.unesp.br/handle/unesp/155306. Acesso em: 02 abr. 2021.

DEY, A. K. Understanding and using context. **Personal and ubiquitous computing**, v. 5, n. 1, p. 4-7, fev. 2001. DOI: https://doi.org/10.1007/s007790170019. Disponível em: https://link.springer.com/article/10.1007/s007790170019. Acesso em: 23 mar. 2021.

DEY, L. *et al*. Ontology aided query expansion for retrieving relevant texts. *In*: SZCZEPANIAK, P.S.; KACPRZYK J.; NIEWIADOMSKI A. (ed). **Advances in Web Intelligence**: Third International Atlantic Web Intelligence Conference, AWIC 2005, Lodz, Poland, June 2005, Proceedings, Berlin: Springer, 2005. p. 126-132.

DICIO. **Dicionário Online de Português**. c.2009-2021. Disponível em: https://www.dicio.com.br/. Acesso em: 28 out. 2019.

DOURISH, P. What we talk about when we talk about context. **Personal and ubiquitous computing**, United Kingdom, v. 8, n. 1, p. 19-30, fev. 2004. DOI: https://doi.org/10.1007/s00779-003-0253-8. Disponível em: https://link.springer.com/article/10.1007/s00779-003-0253-8. Acesso em: 23 mar. 2021.

DUCHARME, B. **Learning SPARQL**: querying and updating with SPARQL 1.1. 2.ed. Sebastopol (USA): O'Reilly Media, Inc., 2013. 366p.

EFTHIMIADIS, E. N. Query expansion. *In*: WILLIAMS, M. E. **Annual review of information science and technology-ARIST**. Medford, N.J.: Information Today, 1996.

ELASTIC NV. **Wikipedia**, 2018. Disponível em: https://en.wikipedia.org/wiki/Elastic_NV. Acesso em: 29 jun. 2020.

ELASTIC. **Elastic.co**, 2018. Disponível em: https://www.elastic.co/pt/. Acesso em: 16 jul. 2020.

ELASTIC, Elasticsearch-PHP. **Elastic.co**, 2020. Disponível em: https://www.elastic.co/guide/en/elasticsearch/client/php-api/current/overview.html. Acesso em: 16 jul. 2020.

FACHIN, G. R. B. Recuperação inteligente da informação e ontologias: um levantamento na área da ciência da informação. **BIBLOS: Revista do Instituto de Ciências Humanas e da Informação**, v. 23, n. 1, p. 259-283, 2009.

FARINELLI, F.; ALMEIDA, M. B. Interoperabilidade semântica em sistemas de informação de saúde por meio de ontologias formais e informais: um estudo da norma OpenEHR. *In*: CONFERÊNCIA INTERNACIONAL ACESSO ABERTO, PRESERVAÇÃO DIGITAL, INTEROPERABILIDADE, VISIBILIDADE E DADOS CIENTÍFICOS, 2014, Porto Alegre. A**nais** [...]. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2014. p. 1-16. Disponível em: http://mba.eci.ufmg.br/downloads/Biredial2014_144_web.pdf. Acesso em: 16 jul. 2020.

FERNÁNDEZ, M. *et al*. Semantically enhanced Information Retrieval: An ontology-based approach. **Journal of Web Semantics**, v. 9, n. 4, p. 434-452, 2011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1570826810000910. Acesso em: 03 mar. 2021.

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

FERNEDA, E. **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.

FERNEDA, E.; DIAS, G. A. Um método de expansão automática de consulta baseada em ontologia. *In*: XIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 14., 2013, Florianópolis. **Anais** [...]. 2013. Florianópolis: UFSC, 2013. Disponível em: http://200.20.0.78/repositorios/handle/123456789/2505. Acesso em: 16 jul. 2020.

FILETO, R. **Sistemas cliente-servidor**. 2006. 16 slides. Disponível em: https://www.inf.ufsc.br/~r.fileto/Disciplinas/BD-Avancado/Aulas/03-ClienteServidor.pdf. Acesso em: 22 jul. 2020.

FLEURY, M. T. L.; WERLANG, S. R. Pesquisa Aplicada: conceitos e abordagens. **Anuário de Pesquisa: 2016-2017**. São Paulo: Fundação Getúlio Vargas, 2017. Acesso em: 05 abr. 2021 Disponível em: https://pesquisa-eaesp.fgv.br/sites/gvpesquisa.fgv.br/files/anuarios/gvpesquisa_2017_final_13112017.pdf. Acesso em: 02 abr. 2021.

FORSETLUND, L.; BJØRNDAL, A. The potential for research-based information in public health: identifying unrecognised information needs. **BMC Public Health**, v. 1, n. 1, p. 1-8, 2001. DOI: https://doi.org/10.1186/1471-2458-1-1. Disponível em: https://link.springer.com/article/10.1186/1471-2458-1-1. Acesso em: 25 abr. 2021.

FU, G.; JONES, C. B.; ABDELMOTY, A. I. Ontology-based spatial query expansion in information retrieval. *In*: MEERSMAN R., TARI Z. (ed.) **On the move to meaningful internet systems 2005**: CoopIS, DOA, and ODBASE. Heidelberg: Springer, 2005. p. 1466-1482. Disponível em: https://link.springer.com/chapter/10.1007/11575801_33. Acesso em: 28 jun. 2021.

FU, L.; GOH, D. H.-L.; FOO, S. S.-B. CQE: a collaborative querying environment. *In*: PROCEEDINGS OF THE 5TH ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL 5., 2005, New York: NY. **Anais** […]. New York: ACM, 2005. p. 378-378.

FUJITA, M. S. L. A política de indexação para representação e recuperação da informação. *In*: GIL-LEIVA, I.; FUJITA, M. S. L. (Eds.). **Política de indexação**. São Paulo: Cultura Acadêmica, 2012. p. 17-28.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GONÇALVES, L. S. **Sistemas de informações gerenciais**. Curitiba: IESDE, 2011. *E-book.*

GONG, Z.; CHEANG, C. W.; HOU, U. L. Web query expansion by WordNet. *In*: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS (DEXA'05), 16., 2005, Copenhagen - DK, **Anais** […]. Copenhagen: IEEE Computer Society, 2005. p. 166-175.

GORMLEY, C.; TONG, Z. **Elasticsearch the definitive guide:** a distributed real-time search and analytics engine**.** Sebastopol (USA): O'Reilly Media, Inc., 2015. 724p.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Human-Computer Studies,** v.43, n.5-6, p.907-928, 1995.

GUARINO, N. Formal ontology in information systems. *In*: GUARINO, N. (ed.). **Formal ontology in information systems**: proceedings of the first international conference (FOIS'98). Amsterdam: IOS Press; Tokyo: Ohmsha, 1998. p. 3-15.

GUARINO, N.; CARRARA, M.; GIARETTA, P. Formalizing ontological commitment. *In*: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), n. 12, 1994, Seattle – Washington. **Anais** […]. Menlo Park: The AAAI Press, 1994. p. 560-567. Disponível em: https://www.aaai.org/Papers/AAAI/1994/AAAI94-085.pdf. Acesso em: 24 abr. 2021.

GUIZZARDI, G. **Desenvolvimento para e com reuso**: um estudo de caso no domínio de vídeo sob demanda. 2000. 202f. Dissertação (Mestrado) - Universidade Federal do Espírito Santo, Centro Tecnológico, Espírito Santo, 2000. Disponível em: http://www.inf.ufes.br/~gguizzardi/dissertacao_msc.pdf. Acesso em: 03 mar. 2021.

GUIZZARDI, G. On ontology, ontologies, conceptualizations, modeling languages, and (meta) models. *In*: VASILECAS, O.; EDER, J.; Caplinskas, A. (Ed.). **Databases and Information Systems IV**. v. 155. Amsterdam: IOS Press. 2007. p. 18-39.

GUTTERIDGE, C. **Sparqllib.php**. 2012. Disponível em: http://graphite.ecs.soton.ac.uk/sparqllib/. Acesso em: 27 jul. 2020.

HARISPE, S. *et al*. Semantic similarity from natural language and ontology analysis. **Synthesis Lectures on Human Language Technologies**, v. 8, n. 1, p. 1-254, 2015. Disponível em: https://www.morganclaypool.com/doi/abs/10.2200/S00639ED1V01Y201504HLT027. Acesso em: 27 jul. 2020.

HARRIS, S.; SEABORNE, A. **SPARQL 1.1 Query Language**. World Wide Web Consortium (W3C), 2013. Disponível em: https://www.w3.org/TR/sparql11-query/. Acesso em: 23 mar. 2021.

HOGAN, A. **The web of data**. Switzerland: Springer, 2020. *E-book* (689 p.). DOI: https://doi.org/10.1007/978-3-030-51580-5. Disponível em: https://www.springer.com/gp/book/9783030515799. Acesso em: 28 jun. 2021.

HORTÊNCIO-FILHO, F. W. B.; LÓSCIO, B. F. Web Semântica: conceitos e tecnologias. *In*: ALCANTARA, P. (Org.). **II Escola Regional de Computação** - Ceará, Maranhão e Piauí - ERCEMAPI 2009.: SBC, 2009.

HOUAISS, A. VILLAR, M. S. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro: Editora Objetiva, 2009.

HYVÖNEN, E. *et al*. MuseumFinland: finnish museums on the semantic web. **Journal of Web Semantics**, Netherlands. v. 3, n. 2-3, p. 224-241, 2005.

IDEHEN, K. U. **What is a SPARQL Endpoint, and why is it important?** 2018. Disponível em: https://medium.com/virtuoso-blog/what-is-a-sparql-endpoint-and-why-is-it-important-b3c9e6a20a8b. Acesso em: 27 jul. 2020.

INGWERSEN, P. **Information retrieval interaction**. London: Taylor Graham, 1992.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Uso de internet, televisão e celular no brasil**. Brasília: IBGE, c2021. Disponível em: https://educa.ibge.gov.br/jovens/materias-especiais/20787-uso-de-internet-televisao-e-celular-no-brasil.html. Acesso em: 26 jun. 2021.

ISOTANI, S.; BITTENCOURT, I. I. **Dados abertos conectados**. São Paulo: Novatec, 2015.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications**: text retrieval, extraction and categorization. 2.ed. Amsterdam: John Benjamins Publishing, 2007.

JASPER, R.; USCHOLD, M. A Framework for understanding and classifying ontology applications. *In*: PROCEEDINGS OF THE IJCAI-99 WORKSHOP ON ONTOLOGIES AND PROBLEM-SOLVING METHODS (KRR5), 16., 1999, Stockholm - Sweden. **Anais** […]. Stockholm, 1999. p. 01-12. Disponível em: http://people.cs.ksu.edu/~abreed/CIS890/References/uschold99.pdf. Acesso em: 27 jul. 2020.

JOACHIMS, T. *et al*. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. **ACM Transactions on Information Systems (TOIS)**, United States, v. 25, n. 2, p. 1-12, 2007. DOI: https://doi.org/10.1145/1229179.1229181.

JONES, R. *et al*. Generating query substitutions. *In*: PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 15., 2006, Edinburgh - Scotland. **Anais** […]. New York: ACM, 2006. p. 387-396. DOI: https://doi.org/10.1145/1135777.1135835.

JQUERY, **What is jQuery?**, c2021. Disponível em: https://jquery.com/. Acesso em: 20 mar. 2021.

KALIBATIENE, D.; VASILECAS, O. Survey on ontology languages. *In*: GRABIS, J., KIRIKOVA, M. (ed). **Perspectives in Business Informatics Research**: 10th International Conference, BIR 2011, Riga, Latvia, October 6-8, 2011. Heidelberg: Springer, 2011. p. 124-141. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-24511-4_10. Acesso em: 28 jun. 2021.

KARA, S. *et al*. An ontology-based retrieval system using semantic indexing. **Information Systems**, United Kingdom, v. 37, n. 4, p. 294-305, jun. 2012. DOI: http://dx.doi.org/10.1016/j.is.2011.09.004.

KENT, A. *et al*. Machine literature searching VIII. Operational criteria for designing information retrieval systems. **American documentation**, United States, v. 6, n. 2, p. 93-101, abr. 1955. DOI: https://doi.org/10.1002/asi.5090060209.

KLYNE, G.; CARROLL, J. (ed.). **Resource description framework (RDF)**: concepts and abstract syntax. World Wide Web Consortium (W3C), 2004. Disponível em: https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. Acesso em: 04 abr. 2021.

KUHLTHAU, C. **Seeking meaning**: A process approach to library and information services. 2. ed. Westport, CT: Libraries Unlimited, 2003.

KUMAR, P. *et al*. Analysis and comparative exploration of elastic search, mongodb and hadoop big data processing. *In*: PANT, M. *et al*. (ed.). **Soft computing**: theories and applications. Singapore: Springer, 2017. p. 605-615. DOI: https://doi.org/10.1007/978-981-10-5699-4_57. Disponível em: https://link.springer.com/chapter/10.1007/978-981-10-5699-4_57. Acesso em: 28 jun. 2021.

KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero - Revista de Ciência da Informação**, Rio de Janeiro, v. 3, n. 1, fev. 2002. Disponível em: http://www.dgz.org.br/fev02/Art_03.htm. Acesso em: 14 dez. 2019.

LAI, J.; SOH, B. Similarity score for information filtering thresholds. *In*: IEEE INTERNATIONAL SYMPOSIUM ON COMMUNICATIONS AND INFORMATION TECHNOLOGY (ISCIT), 2004, Sapporo - Japan. **Anais** […]. New York: IEEE, 2004. p. 216-221. DOI: https://doi.org/10.1109/ISCIT.2004.1413780.

LANCASTER, F. W. **Evaluation of the MEDLARS demand search service**. Washington: National Library of Medicine, 1968.

LE COADIC, Y-F. **A ciência da informação**. 2. ed. Brasilia: Briquet de Lemos, 2004.

LEI, Y.; UREN, V.; MOTTA, E. SemSearch: a search engine for the semantic web. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT, 15., 2006, Podebrady - Czech Republic. **Anais** […]. Heidelberg- Berlim: Springer, 2006. p. 238-245. DOI: https://doi.org/10.1007/11891451_22.

LIMA, V. M. A. **Da classificação do conhecimento científico aos sistemas de recuperação de informação**: enunciação de codificação e enunciação de decodificação da informação documentária. 2004. Tese (Doutorado) - Universidade de São Paulo, Escola de Comunicações e Artes, São Paulo, 2004. Disponível em: https://teses.usp.br/teses/disponiveis/27/27143/tde-06032006-150120/pt-br.php. Acesso em: 23 mar. 2021.

LIMA, J. C.; CARVALHO, C. L. **Uma visão da web semântica**. 2004. 16f. Relatório Técnico - Universidade Federal de Goiás, Instituto de Informática, 2004. Disponível em: https://www.researchgate.net/profile/Junio-Lima/publication/266458321_Uma_Visao_da_Web_Semantica/links/54c23b3a0cf219bbe4e64f1f/Uma-Visao-da-Web-Semantica.pdf. Acesso em: 16 jul. 2020.

LIU, K.; HOGAN, W. R.; CROWLEY, R. S. Natural language processing methods and systems for biomedical ontology learning. **Journal of biomedical informatics**, United States, v. 44, n. 1, p. 163-179, 2011. Disponível em: https://www.sciencedirect.com/science/article/pii/S153204641000105X Acesso em: 16 jul. 2020.

LOPES, L. F. **Um Modelo de engenharia do conhecimento baseado em ontologia e cálculo probabilístico para o apoio ao diagnóstico**. 2011. 233f. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011. Disponível em: http://repositorio.ufsc.br/xmlui/handle/123456789/94742. Acesso em: 03 mar. 2021.

LOURENÇO, A.; RAMALHO, J. C.; PENTEADO, P. Da gestão da informação à gestão do conhecimento: uma proposta para a e-Administração em Portugal. *In*: VII ENCUENTRO IBÉRICO EDICIC 2015. 7., 2015, Madrid. **Anais** [...]. Madrid, 2015. p. 1-13. Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/38496. Acesso em: 03 mar. 2021.

LUBURIĆ, N.; IVANOVIĆ, D. Comparing apache solr and elasticsearch search servers. *In*: INTERNATIONAL CONFERENCE ON INFORMATION SOCIETY AND TECHNOLOGY (ICIST), 6., 2016, Belgrade – Serbia. **Anais** […]. Belgrade: Society for Information Systems and Computer Networks, 2016. p. 287-291. Disponível em: http://www.eventiotic.com/eventiotic/files/Papers/URL/icist2016_54.pdf. Acesso em: 28 jun. 2021.

MAIMONE, G. D.; SILVEIRA, N. C.; TÁLAMO, M. F. G. M. Reflexões acerca das relações entre representação temática e descritiva. **Informação & Sociedade**: Estudos, João Pessoa, v.21, n.1, p. 27-35, jan./abr. 2011. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/7367. Acesso em: 01 out. 2019.

MANAF, N. A. A.; BECHHOFER, S.; STEVENS, R. A survey of identiers and labels in OWL ontologies. *In*: PROCEEDINGS OF THE 6TH INTERNATIONAL WORKSHOP ON OWL EXPERIENCES AND DIRECTIONS (OWLED), 7., 2010, San Francisco-CA. **Anais** […]. San Francisco-CA, 2010. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.364.3255&rep=rep1&type=pdf. Acesso em: 03 mar. 2021.

MANDALA, R.; TOKUNAGA, T.; TANAKA, H. Combining multiple evidence from different types of thesaurus for query expansion. *In*: PROCEEDINGS OF THE 22ND

ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999, Berkeley-CA. **Anais** […]. New York: ACM, 1999. p. 191-197. Disponível em: https://dl.acm.org/doi/pdf/10.1145/312624.312677. Acesso em: 03 mar. 2021.

MANGOLD, C. A survey and classification of semantic search approaches. **International Journal of Metadata, Semantics and Ontologies**, United Kingdom, v. 2, n. 1, p. 23-34, set. 2007. DOI: https://doi.org/10.1504/IJMSO.2007.015073.

MARCHIONINI, G. Information-Seeking Strategies of Novices Using a Full-Text Electronic Encyclopedia. **Journal of the American Society for Information Science**, United States, v.40, n.1, p.54-66. 1989. DOI: https://doi.org/10.1002/(SICI)1097-4571(198901)40:1%3C54::AID-ASI6%3E3.0.CO;2-R.

MCGUINNESS, D. L.; HARMELEN, F. V., **OWL Web Ontology Language**, World Wide Web Consortium (W3C), 2004. Disponível em: https://www.w3.org/TR/owl-features/. Acesso em: 18 mar. 20021.

MEDEIROS, M. B. B., Terminologia brasileira em ciência da informação: uma análise. **Ciência da Informação**, Brasília, v. 15, n. 2, p. 135-142, dez. 1986. Disponível em: http://revista.ibict.br/ciinf/article/view/234. Acesso em: 13 fev. 2021.

MICHAELIS DICIONÁRIO BRASILEIRO DA LÍNGUA PORTUGUESA. [versão digital]. c2021. Disponível em: https://michaelis.uol.com.br/moderno-portugues/. Acesso em 28 out. 2019.

MIZOGUCHI, R. Tutorial on ontological engineering: part 3 - Advanced course of ontological engineering. **New Generation Computing**, Japan, v.22, n.2, p. 198-220, 2004. DOI: http://dx.doi.org/10.1007/BF03040960.

MONTIEL-PONSODA, E. *et al*. Style guidelines for naming and labeling ontologies in the multilingual web. *In*: Proceedings of the International Conference on Dublin Core and Metadata Applications (Online), 2011, United States. **Anais** […] United States: Dublin Core Metadata Initiative, 2011. p. 105-115. Disponível em: https://dcpapers.dublincore.org/pubs/article/view/3626. Acesso em: 03 mar. 2021.

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

MOREIRA, W. **A construção de informações documentárias**: aportes da linguística documentária, da terminologia e das ontologias. 2010. 156f. Tese (Doutorado) - Universidade de São Paulo (USP), Escola de Comunicações e Artes, São Paulo, 2010.

MORENO, A.; PÉREZ, C. Reusing the mikrokosmos ontology for concept-based multilingual terminology databases. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE

RESOURCES AND EVALUATION, 2., 2000, Athens - Greece. **Anais** […]. Athens: The National Technical University, 2000. Disponível em: https://www.researchgate.net/profile/Antonio_Ortiz11/publication/267551147_Reusing_the_Mikrokosmos_Ontology_for_Concept-based_Multilingual_Terminology_Databases/links/573c512008aea45ee84182a7/Reusing-the-Mikrokosmos-Ontology-for-Concept-based-Multilingual-Terminology-Databases.pdf. Acesso em: 03 mar. 2021.

MORRIS, R. C. T. Toward a user-centered information service. **Journal of the American Society for Information Science**, United States, v. 45, n. 1, p. 20-30, 1994. DOI: https://doi.org/10.1002/(SICI)1097-4571(199401)45:1%3C20::AID-ASI3%3E3.0.CO;2-N.

MÜLLER, H. M.; KENNY, E. E.; STERNBERG, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. **PLOS Biology**, United States, v. 2, n. 11, 2004. DOI: https://doi.org/10.1371/journal.pbio.0020309.

NAKANO, N. **Princípios do design da informação na curadoria digital de ambientes virtuais de aprendizagem sob a perspectiva da ciência da informação**. 2019. 165 f. Tese (Doutorado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019.

NASCIMENTO, F. M. S.; PINHO, F. A. Ontologia na gestão de conhecimento jurídico. **Revista P2P & INOVAÇÃO**, Rio de Janeiro, v. 4, n. 2, p. 41-52, mar. 2018. DOI: https://doi.org/10.21721/p2p.2018v4n2.p41-52.

NAVIGLI, R.; VELARDI, P. An analysis of ontology-based query expansion strategies. *In*: PROCEEDINGS OF THE 14TH EUROPEAN CONFERENCE ON MACHINE LEARNING, WORKSHOP ON ADAPTIVE TEXT EXTRACTION AND MINING, 14., 2003, Dubrovnik – Croatia. **Anais** […]. Dubrovnik, 2003. p. 42-49.

NAYAK, T.; PRASAD, S.; SENAPATI, M. R. A survey on web text information retrieval in text mining. **Research Journal of Applied Sciences, Engineering and Technology**, Taiwan, v. 10, n. 10, p. 1164-1174, 2015.

NILSSON, K.; HJELM, H.; OXHAMMAR, H. SUiS–cross-language ontology-driven information retrieval in a restricted domain. *In*: PROCEEDINGS OF THE 15TH NORDIC CONFERENCE OF COMPUTATIONAL LINGUISTICS (NODALIDA 2005), 15., 2006. Joensuu-Finland. **Anais** […] Finland: University of Joensuu, 2006. p. 139-145.

NODINE, M. H.; FOWLER, J. On the Impact of Ontological Commitment. *In*: TAMMA V. *et al*. (ed.). **Ontologies for Agents**: Theory and Experiences. Basel – Switzerland: Springer. 2005. p. 19-42. DOI: https://doi.org/10.1007/3-7643-7361-X_2.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101**: a guide to creating your first ontology. California: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05,

Stanford Medical Informatics Technical Report SMI-2001-0880, 2001. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.5085&rep=rep1&type=pdf. Acesso em: 16 nov. 2019.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.BR). **Pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras**: Pesquisa TIC Educação – 2019, 2020. Disponível em: https://cetic.br/pt/tics/educacao/2019/escolas-urbanas-alunos/C4C/. Acesso em: 26 jun. 2021.

OWL WORKING GROUP, **Web Ontology Language (OWL)**, World Wide Web Consortium (W3C), 2012. Disponível em: https://www.w3.org/OWL/. Acesso em: 11 mar. 2021.

OXFORD LANGUAGES. **Oxford languages and Google**. 2021. Disponível em: https://languages.oup.com/google-dictionary-pt/. Acesso em: 23 mar. 2021.

PABÓN, O. S.; GONZÁLEZ, M. E. S. M. Propuesta para extender semánticamente el proceso de recuperación de información. **Revista EIA**, Colombia, v. 11, n. 22, p. 51-65, jul.-dez. 2014. Disponível em: https://www.redalyc.org/articulo.oa?id=149237906005 Acesso em: 27 jul. 2020.

PANSANI-JUNIOR, E. A. **Ontologias no processo de indexação automática de documentos textuais**. 2016. 127 f. Dissertação (Mestrado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2016.

PEREIRA, C. C. (org.). **Gramática essencial**. Rio de Janeiro: Lexikon, 2013.

PEREIRA, F. R.; RIGO, S. J. Utilização de processamento de linguagem natural e ontologias na análise qualitativa de frases curtas. **RENOTE-Revista Novas Tecnologias na Educação**, Porto Alegre, v. 11, n. 3, 2013. DOI: https://doi.org/10.22456/1679-1916.44431. Disponível em: https://www.seer.ufrgs.br/renote/article/view/44431. Acesso em: 03 mar. 2021.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of SPARQL. **ACM Transactions on Database Systems (TODS)**, v. 34, n. 3, p. 1-45, set. 2009. DOI: https://doi.org/10.1145/1567274.1567278.

PICKLER, M. E. V. **Diretrizes para utilização de ontologias na indexação automática**. 2014. 101f. Dissertação (Mestrado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.

POTTKER, L. M. V. **Arquitetura para recuperação de objetos de aprendizagem**: uma abordagem baseada em agentes inteligentes e Relevance Feedback. 2017. 204 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, Marília, 2017.

PRIBERAM. **Dicionário Priberam da Língua Portuguesa (DPLP)**. c2021. Disponível em: https://dicionario.priberam.org/. Acesso em: 28 out. 2019.

PRUD'HOMMEAUX, E.; SEABORNE, A. **SPARQL Query Language for RDF**. World Wide Web Consortium (W3C), 2008. Disponível em: https://www.w3.org/TR/rdf-sparql-query/. Acesso em: 23 mar. 2021.

PRADO, S. G. D. **Um experimento no uso de ontologias para reforço da aprendizagem em educação à distância**. 2005. 190 f. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2005.

RAMALHO, R.A.S. **Desenvolvimento e utilização de ontologias em bibliotecas digitais**: uma proposta de aplicação. 2010. 145 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, Marília, 2010.

RAZA, M. A. *et al*. A taxonomy and survey of semantic approaches for query expansion. **IEEE Access**, United States, v. 7, p. 17823-17833, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2894679. Disponível em: https://ieeexplore.ieee.org/abstract/document/8625452. Acesso em: 28 jun. 2021.

REDMOND-NEAL, A.; HLAVA, M. M. K. (Ed.). **ASIS&T thesaurus of information science, technology, and librarianship**. Medford: Information Today Inc., 2005.

RIECKEN, R. F. Frame de temas potenciais de pesquisa em ciência da informação. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 3, n. 2, p. 43-63, jan./jun., 2006. DOI: https://doi.org/10.20396/rdbci.v3i2.2044. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/2044. Acesso em: 03 mar. 2021.

RIEH, S. Y.; XIE, H. Patterns and sequences of multiple query reformulations in web searching: A preliminary study. *In*: PROCEEDINGS OF THE ANNUAL MEETING-AMERICAN SOCIETY FOR INFORMATION SCIENCE, 64., 2001, Washington, DC. **Anais** […]. Washington, DC: Information Today, 2001. p. 246-255.

RIJSBERGEN, C. J. V. **Information retrieval**. Glasgow, Scotland, UK: University of Glasgow, 1979.

RINALDI, A. M. An ontology-driven approach for semantic information retrieval on the web. **ACM Transactions on Internet Technology (TOIT)**, New York, v. 9, n. 3, p. 1-24, 2009. DOI: https://doi.org/10.1145/1552291.1552293. Disponível em: https://dl.acm.org/doi/abs/10.1145/1552291.1552293. Acesso em: 24 abr. 2021.

ROA-MARTÍNEZ, S. M. **Da information findability à image findability**: aportes da polirrepresentação, recuperação e comportamento de busca. 2019. 235f Tese (Doutorado) -

Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, Marília, 2019. Disponível em: https://repositorio.unesp.br/handle/11449/182465. Acesso em: 02 out. 2019.

SACK, H. NPBibSearch-An ontology augmented bibliographic search. *In*: PROCEEDINGS 2ND ITALIAN SEMANTIC WEB WORKSHOP, 2., 2005, Trento-Italy. **Anais** […]. Italy: University of Trento, 2005. Disponível em: http://ceur-ws.org/Vol-166/28.pdf. Acesso em: 03 mar. 2021.

SALTON, G. **Automatic information organization and retrieval**. Nova York: McGraw Hill Text, 1968.

SALTON, G. **The SMART retrieval system** – experiments in automatic document processing. Nova Jersey: Prentice Hall. 1971.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**, New York: McGraw-Hill Book Company, 1983.

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

SANTARÉM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação & Informação**, Londrina, v. 21, n. 2, p. 217-244, 2016. DOI: http://dx.doi.org/10.5433/1981-8920.2016v21n2p217. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/26417. Acesso em: 28 jun. 2021.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p .41-62, jan./jun. 1996. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em: 12 mar. 2021.

SCHIESSL, M. **Lexicalização de ontologias**: o relacionamento entre conteúdo e significado no contexto da recuperação da informação. 2015. 261f. Tese (Doutorado) - Universidade de Brasília, Brasília, 2015.

SEVERINO, A. J. **Metodologia do trabalho científico**. 23. ed. São Paulo: Cortez, 2007.

SHAH, N.; WILLICK, D.; MAGO, V. A framework for social media data analytics using elasticsearch and kibana. **Wireless networks**, Switzerland, p. 1-9, 2018. DOI: https://doi.org/10.1007/s11276-018-01896-2. Disponível em: https://link.springer.com/article/10.1007/s11276-018-01896-2. Acesso em: 28 jun. 2021.

SHNEIDERMAN, B. *et al*. **Designing the user interface**: strategies for effective human-computer interaction. 6. ed. Hoboken: Pearson, 2017.

SILVA, A. R. **Aprimorando a visualização e composição de regras SWRL na web**. 2012. 129f. Dissertação (Mestrado) – Universidade de São Paulo, Programa de Pós-Graduação em Matemática Computacional, São Carlos, 2012. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-27022012-142801/pt-br.php. Acesso em: 28 jun. 2021.

SILVA, E. H. **Introdução à programação de computadores**. Santa Catarina, 2017. Disponível em: https://bit.ly/3ztshUD. Acesso em: 28 jun. 2021.

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: 23 mar. 2021. DOI: http://dx.doi.org/10.5433/1981-8920.2013v18n3p27.

SIMOES, M. G. M. *et al*. Indexação automática e ontologias: identificação dos contributos convergentes na ciência da informação. **Ciência Da Informação**, v. 46, n. 1, p.152-168, jan./abr. 2017. Disponível em: http://revista.ibict.br/ciinf/article/view/4020. Acesso em: 23 mar. 2021.

SLIMANI, T. Ontology development: a comparing study on tools, languages and formalisms. **Indian Journal of Science and Technology**, Chennai, v. 8, n. 24, p. 1-12, 2015. DOI: 10.17485/ijst/2015/v8i34/54249. Disponível em: https://www.researchgate.net/profile/Thabet_Slimani1/publication/288258366_Ontology_Development_A_Comparing_Study_on_Tools_Languages_and_Formalisms/links/56d1369608ae85c823487dc6.pdf. Acesso em: 12 nov. 2019.

SMITH, M. K.; WELTY, C.; MCGUINNESS, D. L. OWL Web Ontology Language Guide. World Wide Web Consortium (W3C), 2004. Disponível em: https://www.w3.org/TR/2004/REC-owl-guide-20040210/. Acesso em: 27 jul. 2020.

SONNENWALD, D. H.; WILDEMUTH, B. S.; HARMON, G. L. A Research Method to Investigate Information Seeking Using the Concept of Information Horizons: An Example from a Study of Lower Socio-Economic Students' Information Seeking Behaviour. **The New Review of Information Behaviour Research**, London, v.2, p.65-86. 2001. Disponível em: http://hdl.handle.net/10760/7969. Acesso em: 03 mar. 2021.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 2, p. 161-173, 2006.

SOWA, J. F. Architectures for intelligent systems. **IBM Systems Journal**, United States, v. 41, n. 3, p. 331-349, 2002. Disponível em: https://ieeexplore.ieee.org/abstract/document/5386878. Acesso em: 14 nov. 2019.

SOWA, J. F. **Conceptual structures**: information processing in mind and machine. Massachussetts: Addison-Wesley Longman Publishing Co., 1984.

SPARK-JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, United Kingdom, v. 28, n. 1, p. 11-21, 1972. DOI: https://doi.org/10.1108/eb026526.

SPARQL WORKING GROUP, **SPARQL Query Language for RDF**. 2008. Disponível em: https://www.w3.org/TR/rdf-sparql-query/. Acesso em: 27 jul. 2020.

SPINK, A.; SARACEVIC, T. Dynamics of search term selection during mediated online searching. *In*: PROCEEDINGS OF THE ASIS ANNUAL MEETING, 30., 1993. Columbus. **Anais** […]. New York: American Society for Information Science, 1993. p. 63-72.

SPINK, A. *et al*. Searching the web: The public and their queries. **Journal of the American society for information science and technology**, v. 52, n. 3, p. 226-234, 2001.

STATCOUNTER GLOBAL STATS. **Search engine market share worldwide**. c1999-2021. Disponível em: https://gs.statcounter.com/search-engine-market-share. Acesso em: 28 jun. 2021.

STOJANOVIC, N. On the query refinement in the ontology-based searching for information. **Information Systems**, United Kingdom, v. 30, n. 7, p. 543-563, 2005. DOI: https://doi.org/10.1016/j.is.2004.11.004.

SULTAGUBIYEVA, A.; RAUSHANGUL, A.; GULZHAMAL, K. Concept “Heart” in the Language Picture of World. **International Journal of Humanities Social Sciences and Education (IJHSSE)**, Ongole, v. 2, n. 1, p. 116-120, 2015. Disponível em: https://www.arcjournals.org/pdfs/ijhsse/v2-i1/15.pdf Acesso em 21 abr. 2021

TANENBAUM, A. S. **Redes de computadores**. 4. ed. Porto Alegre: Bookman, 2003.

TAYLOR, R. S. Question-Negotiation and Information Seeking in Libraries. **College & Research Libraries**, v. 76, n. 3, p. 251-267, mar. 2015. DOI: https://doi.org/10.5860/crl.76.3.251. Disponível em: https://crl.acrl.org/index.php/crl/article/view/16421/17867. Acesso em: 24 abr. 2021.

THIOLLENT, M. **Metodologia da pesquisa-ação**. 18. ed. São Paulo: Cortez Editora, 2018.

THOMÉ, M. **Como funcionam os mecanismos de busca**, João Pessoa, 2017. Disponível em: https://administradores.com.br/artigos/como-funcionam-os-mecanismos-de-busca. Acesso em: 15 mar. 2021.

TORRES, S.; ALMEIDA, M. B. O conceito de documento na ciência da informação e arquivologia. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA

INFORMAÇÃO (ENANCIB), 14., 2013, Florianópolis. **Anais** [...]. Florianópolis: UFSC, ANCIB, 2013. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2417. Acesso em: 28 jun. 2021.

TRAMONTIN JUNIOR, R. J. **Um Modelo baseado em contexto para expansão de consultas semânticas em redes colaborativas de organizações**. 2011. 309 f. Tese (Doutorado) – Universidade Federal de Santa Catarina, Santa Catarina, 2011.

USCHOLD, M.; GRUNINGER, M. Ontologies: Principles, methods and applications. **The knowledge engineering review**, Cambridge, v. 11, n. 2, p. 93-136, 1996. Disponível em: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/ontologies-principles-methods-and-applications/2443E0A8E5D81A144D8C611EF20043E6. Acesso em: 03 mar. 2021.

VALLET, D.; FERNÁNDEZ, M.; CASTELLS, P. An ontology-based information retrieval model. *In*: GÓMEZ-PÉREZ A., EUZENAT J. (eds) **The Semantic Web**: research and applications. Crete-Greece: Springer, 2005. p. 455-470. DOI: https://doi.org/10.1007/11431053_31.

VANOYE, F. **Usos da linguagem**: problemas e técnicas na produção oral e escrita. São Paulo: Martins Fontes, 1992.

VICKERY, B. C. On ‘knowledge organisation’. *In*: GILCHRIST, A.; VERNAU, J. (ed.). **Facets of Knowledge Organization**: proceedings of the ISKO UK Second Biennial Conference, Bingley-UK: Emerald Group Publishing, 2012. p. 203-232.

VIEIRA, S. B. **La recuperación automática de información jurídica**: metodología de análisis lógico-sintáctico para la lengua portuguesa. 1994. 402f. Tese (Doutorado) - Universidad Complutense de Madrid, Madrid, 1994.

VIJAYARAJAN, V. *et al*. A generic framework for ontology-based information retrieval and image retrieval in web data. **Human-centric Computing and Information Sciences**, Heidelberg – Germany, v. 6, n. 1, p. 1-30, 2016. DOI: https://doi.org/10.1186/s13673-016-0074-1. Disponível em: https://hcis-journal.springeropen.com/articles/10.1186/s13673-016-0074-1. Acesso em: 24 abr. 2021.

VOIT, A. *et al*. Big data processing for full-text search and visualization with elasticsearch. **International journal of advanced computer science and applications**, United States, v. 8, n. 12, p. 76-83, 2017. DOI: http://dx.doi.org/10.14569/IJACSA.2017.081211. Disponível em: https://thesai.org/Publications/ViewPaper?Volume=8&Issue=12&Code=ijacsa&SerialNo=11. Acesso em: 28 jun. 2021.

WORLD WIDE WEB CONSORTIUM (W3C). **About the world wide web**. 1992. Disponível em: https://www.w3.org/WWW/. Acesso em: 17 jul. 2020.

WORLD WIDE WEB CONSORTIUM (W3C). **HTML & CSS**. 2016. Disponível em: https://www.w3.org/standards/webdesign/htmlcss. Acesso em:17 jul. 2020.

WORLD WIDE WEB CONSORTIUM (W3C). **W3C Semantic web activity**. 2013. Disponível em: https://www.w3.org/2001/sw/. Acesso em: 23 mar. 2021.

WORLD WIDE WEB TECHNOLOGY SURVEYS (W3TECHS). **Usage statistics of PHP for websites**. c2009-2021. Disponível em: https://w3techs.com/technologies/details/pl-php. Acesso em: 17 jul. 2020.

WATSON, M. **Practical Semantic Web and Linked Data Applications**. Mark Watson, 2010. *E-book* (190 p.).

WEI, W.; BARNAGHI, P. M.; BARGIELA, A. Search with meanings: an overview of semantic search systems. **International journal of Communications of SIWN**, v. 3, p. 76-82, 2008.

WILLARD VAN ORMAN QUINE. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation, 2021]. Disponível em: https://en.wikipedia.org/w/index.php?title=Willard_Van_Orman_Quine&oldid=1031739424. Acesso em: 24 abr. 2021.

WILLIAMSON, J. W. *et al*. Health science information management and continuing education of physicians: a survey of US primary care practitioners and their opinion leaders. **Annals of Internal Medicine**, v. 110, n. 2, p. 151-160, 1989. DOI: https://doi.org/10.7326/0003-4819-110-2-151. Disponível em: https://www.acpjournals.org/doi/abs/10.7326/0003-4819-110-2-151. Acesso em: 25 abr. 2021.

WOODS, W. A. **Searching versus finding**: why systems need knowledge to find what you really want. Sun Microsystems Laboratories, 2004. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.595&rep=rep1&type=pdf. Acesso em: 23 mar. 2021.

WORDNET. **What is WordNet?**. c2021. Disponível em: https://wordnet.princeton.edu/. Acesso em: 08 jun. 2020.

XIANG, B. *et al*. Context-aware ranking in web search. *In*: PROCEEDING OF THE 33RD INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 33., 2010, Geneva – Switzerland. **Anais** […]. New York: ACM, 2010. p. 451-458.

YUNZHI, C. *et al*. An approach to semantic query expansion system based on hepatitis ontology. **Journal of Biological Research**, Thessaloniki - Greece, v. 23, n. 1, p. 11-22, 2016. DOI: https://doi.org/10.1186/s40709-016-0044-9. Disponível em: https://link.springer.com/article/10.1186/s40709-016-0044-9. Acesso em: 28 jun. 2021.

ZHANG, J.; DENG, B.; LI, X. Concept based query expansion using wordnet. *In*: INTERNATIONAL E-CONFERENCE ON ADVANCED SCIENCE AND TECHNOLOGY, 2009, Washington – DC. **Anais** […]. Washington: IEEE, 2009. p. 52-55.