UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Filosofia e Ciências
Campus Marília
Programa de Pós-Graduação em Ciência da Informação

**CAIO SARAIVA CONEGLIAN**

# Recuperação da Informação com abordagem semântica utilizando Linguagem Natural: a Inteligência Artificial na Ciência da Informação

**Marília – SP**

Universidade Estadual Paulista “Júlio Mesquita Filho”
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

CAIO SARAIVA CONEGLIAN

# Recuperação da Informação com abordagem semântica utilizando Linguagem Natural: a Inteligência Artificial na Ciência da Informação

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Estadual Paulista – Campus de Marília, como requisito para a obtenção do título de doutor em Ciência da Informação.

**Área de Concentração**: Informação, Tecnologia e Conhecimento.
**Linha de Pesquisa**: Informação e Tecnologia
**Financiamento**: CAPES e CNPQ
**Orientador**: Dr. José Eduardo Santarem Segundo

Marília – SP
2020

Coneglian, Caio Saraiva
C747r Recuperação da Informação com abordagem semântica utilizando Linguagem Natural : a Inteligência Artificial na Ciência da Informação / Caio Saraiva Coneglian. -- Marília, 2020
195 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília
Orientador: José Eduardo Santarém Segundo

1. Recuperação da informação. 2. Web Semântica. 3. Inteligência artificial. 4. Processamento de linguagem natural (Computação). 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.

Caio Saraiva Coneglian

# RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA UTILIZANDO LINGUAGEM NATURAL: A INTELIGÊNCIA ARTIFICIAL NA CIÊNCIA DA INFORMAÇÃO

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

Área de Concentração: Informação, Tecnologia e Conhecimento.
Linha de Pesquisa: Informação e Tecnologia
Data da defesa: 03 de julho de 2020
Local: Faculdade de Filosofia e Ciências, UNESP - Campus de Marília

## BANCA EXAMINADORA

**Presidente e Orientador: Prof. Dr. José Eduardo Santarem Segundo**
Professor no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília. Professor da Universidade de São Paulo, USP. Campus de Ribeirão Preto.

**Membro Titular: Profa. Dra. Silvana Aparecida Borsetti Gregorio Vidotti**
Professora no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília

**Membro Titular: Prof. Dr. Edberto Ferneda**
Professor no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP Campus de Marília

**Membro Titular Externo: Prof. Dr. Elvis Fusco**
Professor da Escola Digital do Centro Universitário Eurípides de Marília, UNIVEM

M**embro Titular Externo: Profa. Dra. Sandra Milena Roa Martinez**
Professora da Universidad del Cauca, UNICAUCA

*Dedico à Natalia, à Ana Maria e à José Artur.*

# AGRADECIMENTOS

À Natalia, por todo o amor, apoio e companheirismo, durante todos estes anos. Agradeço por estar presente e me ajudar no desenvolvimento deste trabalho e apoio em todos estes momentos.

Aos meus pais, Ana Maria e José Artur, que durante meu doutorado e minha vida me apoiaram e possibilitaram a realização deste trabalho, me inspirando a crescer a cada dia.

Aos meus familiares, Fernando, Seforah, Maria Fernanda e Rafael, que estiveram presentes em diversos momentos e sempre me apoiaram e trouxeram mais alegrias para as nossas vidas.

Ao professor José Eduardo Santarem Segundo, que me orientou durante toda a pós-graduação, me ensinando sobre a Ciência da Informação, os métodos científicos e questões éticas e morais, tendo a minha admiração por toda a sua trajetória e conhecimento.

Ao Prof. Elvis Fusco, que desde a minha graduação tem me auxiliado e ensinado a ser profissional, cientista e ser crítico com todas as coisas. Sou muito grato por todas as oportunidades e ensinamentos me passados, sendo parte essencial em minha trajetória profissional e acadêmica.

À Profa. Silvana Aparecida Borsetti Gregorio Vidotti, que me trouxe grandes ensinamentos sobre a Ciência da Informação e a vida. Agradeço imensamente a oportunidade de ter convivido e aprendido com uma das grandes referências da área no país.

À Profa. Sandra Milena Roa-Martinez, que se tornou uma grande amiga e companheira durante o meu doutorado, me ensinando muito sobre como podemos ser mais críticos e exigentes com as coisas. Sou grato por ser parte da minha banca e minha amiga.

Ao Prof. Edberto Ferneda, por ter aceito o convite de ser parte da minha banca e contribuir com os seus conhecimentos.

Aos meus grandes amigos, Felipe e Luís Fernando, que estiveram presentes há muitos anos na minha vida, e me auxiliaram a crescer profissionalmente e pessoalmente. Muito obrigado por serem meus amigos.

Aos meus amigos, Anderson, Carol, Adriana e Lucas, pessoas muito especiais e importantes, que participaram e me apoiaram neste processo.

Às minhas avós, meus tios, primos e sogros, que parte da minha família, estiveram presentes e apoiaram durante o processo.

Aos meus amigos e colegas da Unesp, em especial, a Ana Maria, Felipe, Paula e Manu, que participaram ativamente no processo de definição e construção da minha tese. Em especial,

Ana Maria e Paula, que me ensinaram muito e foram fundamentais para a minha inserção na área de Ciência da Informação.

Aos meus amigos e colegas do UNIVEM, em especial, Fábio Dacêncio Pereira, Rodolfo Chiaramonte e Fabio Piola Navarro, que me acolheram e apoiaram neste processo desafiador e fundamental na minha carreira profissional.

Aos demais professores da Unesp, em especial, Leonardo Botega, Rachel Alves e Marta Valentim, pelo aprendizado a mim possibilitado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Às agências de fomento que me apoiaram, CAPES e CNPq, que financiaram parte do meu doutoramento.

Ao UNIVEM, pelo apoio durante este período.

À Profa. Maria Lúcia Balestriero, pela correção ortográfica.

À Deus, por tudo.

# RECUPERAÇÃO DA INFORMAÇÃO COM ABORDAGEM SEMÂNTICA UTILIZANDO LINGUAGEM NATURAL: A INTELIGÊNCIA ARTIFICIAL NA CIÊNCIA DA INFORMAÇÃO


## RESUMO

A evolução das Tecnologias de Informação e Comunicação conduziu ao desenvolvimento de técnicas capazes de recuperar informações com mais eficiência, inclusive aproximando a linguagem computacional da linguagem natural. Nesse sentido, técnicas de recuperação da informação que utilizam processamento de linguagem natural, como o *Question Answering*, e a Web Semântica, podem ser utilizados em conjunto para aprimorar a satisfação das necessidades informacionais dos usuários. No âmbito da Web Semântica, as ontologias e o *Linked Data* podem ser utilizados como uma importante fonte informacional, por contemplar conhecimentos de diversas áreas do conhecimento. Somado a esse cenário, há a dificuldade eminente dos usuários utilizarem sistemas de recuperação da informação que não levam em consideração a sua linguagem natural, tampouco a semântica dos termos de busca e o contexto dos dados das fontes informacionais. Dessa forma, esta pesquisa apresenta como objetivo a proposição de um modelo de recuperação da informação que redesenha este campo de estudos, a partir da aproximação da linguagem computacional com a linguagem natural, utilizando os princípios da representação da informação, para que o significado e o contexto dos dados estejam explícitos para o processo da busca; para tanto, aproxima-se e relaciona-se aos processos de Inteligência Artificial, processamento de linguagem natural e às ferramentas da Web Semântica. Para o desenvolvimento deste trabalho, utilizou-se o método quadripolar, sendo um estudo exploratório e aplicado. Como resultados, criou-se este modelo de recuperação da informação, pautado no contexto semântico e na aplicação da Inteligência Artificial, capaz de tornar a linguagem natural a base do processo, e considerando o contexto e o significado dos termos para os usuários. Aponta-se que tal modelo é capaz de aprimorar a satisfação das necessidades informacionais dos usuários, utilizando as ontologias para contextualizar as informações, o *Linked Data* para fornecer dados estruturados e o processamento de linguagem natural para aproximar a linguagem computacional da linguagem natural. Outro resultado está na prova de conceito, que demonstra a validade e a aplicação do modelo, apresentando um caso real de como o processo de recuperação da informação ocorre neste modelo, com todas as possibilidades e como as diversas ferramentas, conceitos e tecnologias se vinculam e promovem o processo na prática. Conclui-se que um modelo de recuperação da informação, quando se utiliza da linguagem natural como padrão, quando apoiado pela Web Semântica e aprendizagem de máquina, torna o processo mais natural, eficaz e acessível, de forma que qualquer usuário será capaz de se utilizar dele, mesmo que não tenha domínio dos mecanismos de busca e recuperação. Além disso, aponta-se que o presente trabalho realiza uma importante aproximação entre a Ciência da Informação e a Inteligência Artificial, trazendo para seu escopo, em especial no âmbito da recuperação da informação, aplicações reais de como este segundo campo de estudos pode aprimorar a área como um todo.



## PALAVRAS-CHAVE:

Recuperação da informação; Web Semântica; Inteligência artificial; Processamento de linguagem natural; Ontologia; Aprendizagem de máquina.

# INFORMATION RETRIEVAL WITH A SEMANTIC APPROACH USING NATURAL LANGUAGE: ARTIFICIAL INTELLIGENCE IN INFORMATION SCIENCE


## ABSTRACT

The evolution of Information and Communication Technologies has led to the development of techniques capable of retrieve information more efficiently, including bringing computational language closer to natural language. In this sense, information retrieval techniques that use Natural Language Processing, such as Question Answering, and the Semantic Web, can be used together to improve the satisfaction of users' information needs. Within the scope of the Semantic Web, ontologies and Linked Data can be used as an important information source, as it contemplates knowledge from different areas of knowledge. In addition to this scenario, there is an eminent difficulty for users to use information retrieval systems that do not consider their natural language, nor the semantics of search terms and the context of data from information sources. In this way, the present research has as objective the proposition of a model of Information Retrieval, that redraws this field of studies, from the approximation of the computational language with the natural language, using the principles of the representation of the information so that the meaning and the context of the data are explicit for the search process, for this purpose the Artificial Intelligence, Natural Language Processing and the Semantic Web tools are related and related. For the development of this work, the quadripolar method was used, being an exploratory and applied study. As results, this Information Retrieval model was created, based on the semantic context and the application of Artificial Intelligence, capable of making natural language the basis of the process, and considering the context and the meaning of the terms for users. It is pointed out that such a model can improve the satisfaction of users' informational needs, using ontologies to contextualize information, Linked Data to provide structured data and Natural Language Processing to bring computational language closer to natural language. Another result is in the proof of concept, which demonstrates the validity and application of the model, presenting a real case of how the information retrieval process occurs in this model, with all the possibilities and how the different tools, concepts and technologies are linked and promote the process in practice. We conclude that an Information Retrieval model, which when using natural language as a standard, when supported by the Semantic Web and machine learning, makes the process more natural, effective and accessible, since any user will be able to use, even if he has no control over search and retrieval mechanisms. Furthermore, it is pointed out that the present work makes an important approximation between Information Science and Artificial Intelligence, bringing to its scope, especially in the scope of Information Retrieval, real applications of how this second field of studies can improve the area as a whole.



## KEYWORDS:

Information Retrieval; Semantic Web; Artificial intelligence; Natural Language Processing; Ontology; Machine learning.

# LISTA DE ILUSTRAÇÕES

# LISTA DE QUADROS

# SUMÁRIO

# 1 INTRODUÇÃO

A Ciência da Informação vem passando por transformações em seus estudos, processos, métodos e práticas, acompanhando as mudanças que a sociedade está vivendo. As Tecnologias da Informação e Comunicação (TIC) estão mais presentes e mais consolidadas no cotidiano das pessoas, desafiando pesquisadores a compreenderem esse processo e, mais especificamente na área da Ciência da Informação, a encontrarem meios para aprimorar o trato com a informação, elemento cada vez mais valorizado na sociedade.

Com a expansão e popularização da Inteligência Artificial (IA), há novos desafios a serem considerados pela área de Ciência da Informação, que exigem pesquisas e práticas para possibilitar uma evolução em diversos campos de estudos, em especial da representação e da recuperação da informação.

Destaca-se que a Inteligência Artificial vem tendo aproximações com a área da Ciência da Informação há alguns anos, com pesquisas sendo realizadas desde o início dos anos 1990 (CUNHA, KOBASHI, 1991). No entanto, com a atual evolução das tecnologias, em especial do campo da Inteligência Artificial, há um cenário bastante propício para o aprofundamento destas discussões e de aplicações da IA dentro da Ciência da Informação.

Ao trazer a abordagem de Borko (1968, tradução nossa), a Ciência da Informação é definida como:

> [...] uma ciência interdisciplinar que investiga as propriedades e comportamento da informação, as forças que governam os fluxos e os usos da informação e as técnicas, tanto manual quanto mecânica, de processamento da informação, visando sua armazenagem, recuperação e disseminação total.

Partindo da definição de Borko, é possível identificar que a Ciência da Informação se preocupa com as diversas etapas relacionadas ao tratamento da informação, passando pelo seu armazenamento, recuperação e disseminação. Complementarmente, Saracevic (1995) insere a recuperação da informação como a mais importante atividade da Ciência da Informação, pois, segundo o autor, é nessa atividade que ocorre a maior parte das relações interdisciplinares dessa ciência.

Como relatado, a expansão das TIC, em especial no contexto da sociedade em rede[1], trouxe novos desafios à Ciência da Informação, especialmente na busca de aproximar a

[1] Sociedade em rede é um termo cunhado por Manuel Castells que apresenta a configuração atual da sociedade, em especial pela dinâmica das redes existentes. Castells (2007, p. 565) afirma que estamos em: "[...] uma sociedade que, portanto, podemos apropriadamente chamar de sociedade em rede, caracterizada pela primazia da morfologia social sobre a ação social."

comunicação entre o analógico e o digital, permitindo que os indivíduos sejam capazes de navegar pelo digital e recuperar as informações com êxito, de forma natural.

Nesse sentido, o processo de recuperar informação nos ambientes atuais não pode ser compreendido da mesma forma como era tratado na metade do século XX, com os tradicionais métodos de recuperação da informação, e seguindo modelos estáticos. Em um momento de complementaridade entre analógico e digital, em que, segundo Castells (2015, p. 37), "[...] a sociabilidade é reconstruída [...] em um processo que combina interação on-line com interação off-line, ciberespaço e espaço local.", é necessário repensar e refletir o modo como um sistema de computador recupera e apresenta as informações aos indivíduos.

A Ciência da Informação tem um papel central nesse processo, tanto na aproximação das TIC com os indivíduos, quanto em fornecer teorias e técnicas capazes de auxiliar no desenvolvimento de ferramentas que aprimoram a forma como a recuperação da informação ocorre. Essa atuação se expandiu com o aumento na geração de dados, que a evolução das TIC provocou. Indubitavelmente, a Web é uma das principais responsáveis por esse cenário.

Desde a sua concepção, em 1989, a Web apresentou um grande crescimento, fazendo com que, passados mais de 30 anos, o número de pessoas que se encontram dentro dessa plataforma ultrapasse os três bilhões. Uma consequência é a criação de documentos digitais em números incalculáveis.

No entanto, o princípio da Web marcou uma mudança definitiva na forma como os indivíduos e as organizações interagiam. Consequentemente, a quantidade de documentos disponível nesse ambiente cresceu de forma exponencial, o que tornou a organização e a recuperação de informação no início da Web ineficiente. Souza e Alvarenga (2004, p. 133) expressam os primeiros anos da Web, pontuando que "[...] a Web foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade."

Aspirando resolver tais questões, três pesquisadores de notável importância nos estudos sobre a Web, Berners-Lee, Hendler e Lassila, propuseram em 2001 a chamada Web Semântica. Em síntese, a Web Semântica buscava tornar a Web um local que promovesse de forma eficiente a comunicação entre agentes computacionais e seres humanos. No texto inicial da Web Semântica, os seus criadores afirmam que:

> A **Web Semântica não é uma Web separada, mas uma extensão da atual,** em que a informação tem um significado bem definido, permitindo que computadores e pessoas trabalhem melhor em cooperação. [...] Em um futuro próximo, esse desenvolvimento irá inaugurar novas funcionalidades

> significativas sobre **como as máquinas se tornam muito mais capazes de processar e "entender" os dados do que simplesmente exibi-los**. (BERNERS-LEE; HENDLER; LASSILA, 2001, não paginado, tradução nossa, grifo nosso).

As proposições destacadas explicitam que a Web Semântica não é uma nova Web, mas sim uma busca de solucionar os problemas apresentados, além de enlaçar a Web Semântica à compreensão que as máquinas devem possuir sobre o conteúdo disponível digitalmente. No âmbito da Ciência da Informação, diversos trabalhos foram publicados pela estreita relação existente entre essa área do conhecimento, a Ciência da Informação, e esse campo de estudos, a Web Semântica.

Em um desses estudos, Ramalho, Vidotti e Fujita (2007, não paginado) apontam que a Web Semântica "[...] tem como objetivo desenvolver meios para que as máquinas possam servir aos humanos de maneira mais eficiente, mas para isso torna-se necessário construir instrumentos que forneçam sentido lógico e semântico aos computadores."

O relato dos autores contempla e reafirma a necessidade de fornecer lógica e semântica aos mecanismos computacionais. Outrossim, Santarem Segundo (2012, p. 107) aprofunda as relações existentes entre a apresentação e a estrutura das informações, declarando que o principal ponto da Web Semântica é a: "[...] "separação da apresentação do conteúdo e do conteúdo da estrutura, tratando as unidades atômicas de uma informação como componentes independentes."

A afirmação do autor sintetiza a Web Semântica, ao passo que contempla as características da informação. Nesse sentido, a Ciência da Informação, por meio de seu arcabouço teórico, conduz a Web Semântica a aprofundar as relações com questões fundamentais, como a representação e a organização das informações. Esses tópicos são essenciais para que a Web Semântica possa desempenhar com mais eficiência a sua proposta inicial, possibilitando uma melhor compreensão dos conteúdos por agentes computacionais.

Para tornar real essa proposta, as ontologias representam um papel essencial, pois propiciam um ferramental adequado para contextualizar as informações e prover um modelo de dados capaz de tornar compreensível as informações para os computadores. (BERNERS-LEE; HENDLER; LASSILA, 2001; SANTAREM SEGUNDO; CONEGLIAN, 2015).

Para traçar o papel que as ontologias apresentam dentro da Web Semântica, Santarem Segundo e Coneglian (2015, p. 227) afirmam que: "Para o uso como tecnologia da Web Semântica, entendem-se as ontologias como artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de classes, propriedades, relações, restrições, axiomas e instâncias."

Complementando o papel das ontologias, Santarem Segundo (2015, p. 226) relata que:

> Utilizar ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação.

A caracterização apresentada possibilita verificar a intrínseca relação entre as ontologias e a contextualização de um domínio. Assim, torna-se necessário a criação de tecnologias e ferramentas que tornem implementável esse conceito de ontologia. Nesse sentido, a linguagem *Web Ontology Language* (OWL) vem ganhando destaque, pois é a linguagem para a construção de ontologias recomendada pela *World Wide Web Consortium* (W3C), consórcio responsável por criar padrões para a Web.

A linguagem OWL é definida como "[...] uma linguagem de Web Semântica projetada para representar o conhecimento rico e complexo sobre as coisas, grupos de coisas, e as relações entre as coisas." (WORLD WIDE WEB CONSORTIUM, 2012, não paginado). Compreende-se por coisas, quaisquer elementos do mundo real, de forma que a OWL é capaz de traçar computacionalmente as relações existentes entre os objetos.

Apesar de tecnologias como as ontologias estarem bastante consolidadas, a princípio, houve poucas aplicações que fizeram uso dos conceitos e das ferramentas da Web Semântica de fato. Contudo, mais recentemente, diversas pesquisas e iniciativas começaram a materializar essa proposta, colocando em prática as ideias pensadas pelos criadores Berners-Lee, Hendler e Lassila. A principal iniciativa dessa materialização é o chamado *Linked Data*, que tem como princípio utilizar as tecnologias e as ferramentas da Web Semântica para inserir significado ao interligar os dados, propiciando a descoberta de informações pelos usuários durante a navegação. (BERNERS-LEE, 2006; SANTAREM SEGUNDO, 2015).

O *Linked Data* cria um cenário para a disponibilização de documentos dentro da Web, em que os dados estão interligados, sendo que as relações seguem os princípios da Web Semântica, inserindo significado a tais relações. Vale destacar que um dos princípios do *Linked Data* está vinculado à descoberta de novas informações, possibilitando que os usuários, ao navegar pelas relações, possam localizar informações relacionadas e relevantes.

As iniciativas da Web Semântica, como um todo, têm como princípio aproximar o entendimento que as máquinas apresentam dos conteúdos gerados pelos humanos, para que, assim, possa haver uma melhor recuperação da informação pelos usuários. Indo ao encontro desse desafio, técnicas de processamento de linguagem natural (PLN) estão buscando

aproximar a linguagem computacional da linguagem humana, buscando oferecer formas de transformar o que uma pessoa escreve e fala em algo compreensível para as máquinas.

Nesse contexto, vale apontar que o processamento de linguagem natural é um subcampo da Inteligência Artificial, relativo ao tratamento e compreensão de textos e da fala por mecanismos computacionais. Nas últimas décadas, a Inteligência Artificial tem se destacado por possibilitar a realização de tarefas de difícil automatização, de forma mais simples e rápida. A evolução da Inteligência Artificial conduziu a uma expansão em serviços e soluções que favorecem a adoção dessa tecnologia, impactando em diversas áreas do conhecimento, como medicina, direito, física, entre outras.

Desde as primeiras referências à Inteligência Artificial, nas décadas de 1940 e 1950, esse campo de estudos foi evoluindo de forma significativa, tendo desde uma abordagem mais filosófica, que tratava de questões de cunho mais teórico, passando pela abordagem matemática, com o estudo de algoritmos e probabilidades, pela abordagem psicológica, com um estudo cognitivo, chegando até à abordagem computacional, com enfoque nos artefatos que podem ser construídos, e à linguística, focada nas representações do conhecimento.

Na área da Ciência da Informação, a Inteligência Artificial é capaz de auxiliar diversos processos, como a representação, organização e recuperação da informação, permitindo a proposição de novos modelos que utilizam essas novas ferramentas e conceitos. A popularização desses serviços de Inteligência Artificial pode impactar significativamente nas pesquisas e soluções criadas para essa área e, ao mesmo tempo, é capaz de criar paradigmas no modo como processos tradicionais são concebidos.

Na recuperação da informação, a Inteligência Artificial e a área de processamento de linguagem natural são capazes de transformar o modo como o usuário interage com o processo de busca e de recuperação, favorecendo uma interação mais natural, conduzida por linguagem natural. Nesse sentido, o usuário, ao invés de se adequar à linguagem computacional para realizar a busca, por exemplo, utilizando palavras-chave e expressões como AND e OR, passa a interagir com o sistema por meio de voz ou da escrita, de forma natural.

Essa mudança de paradigma do campo de recuperação da informação é possível devido às tecnologias de Inteligência Artificial que permitem a compreensão do que o usuário está falando ou escrevendo. Além disso, aponta-se que a lacuna existente entre a linguagem computacional e a natural pode ser diminuída com esforços que busquem aprimorar a compreensão dos textos e termos utilizados pelas pessoas. Em especial, o apoio das ferramentas da Web Semântica, junto à Inteligência Artificial, é capaz de aumentar a compreensão da

semântica e da pragmática dos termos, bem como dos conceitos de um texto, favorecendo uma melhor recuperação da informação.

Nesse contexto, uma técnica de recuperação da informação que tem como princípio o PLN, é o chamado *Question Answering* (QA), um sistema que responde de forma objetiva a perguntas feitas por usuários. O uso de QA tem sido bastante difundido por ser uma forma de tornar o processo de busca mais natural, sem que o usuário tenha que se preocupar em escrever palavras-chave abrangentes e fazer uso de técnicas de buscas que exijam um certo conhecimento prévio.

Allam e Haggag (2012) afirmam que os estudos de QA abrangem pesquisas dentro de recuperação da informação, extração da informação e PLN. Os autores complementam relatando que: "[...] o principal objetivo de todos os sistemas de QA é recuperar respostas a perguntas, em vez de documentos completos ou dos melhores trechos desses documentos, como a maioria dos sistemas de recuperação de informação faz atualmente." (ALLAM; HAGGAG, 2012, p. 211, tradução nossa).

A união entre as técnicas de PLN, QA e Inteligência Artificial com o uso do ferramental da Web Semântica, especialmente as ontologias e o *Linked Data*, pode aprimorar o processo de recuperação da informação, ao aproximar o entendimento da linguagem natural, além de permitir a contextualização dos dados durante a recuperação da informação.

Nesse contexto, a Ciência da Informação apresenta função essencial, no que tange à ótica que essa área do conhecimento possui sobre a recuperação da informação, bem como aos estudos pragmáticos e teóricos que a Ciência da Informação vem elaborando acerca da Web Semântica. A união entre a Inteligência Artificial e a Web Semântica realizada na seara da Ciência da Informação pode se aperfeiçoar nos procedimentos realizados, especialmente pelas preocupações existentes na organização, na representação e na descrição da informação.

Vale destacar, ainda, que os estudos com ontologias desenvolvidos dentro da Ciência da Informação conseguem aprofundar o entendimento sobre as propriedades das ontologias, a contextualização dos dados e o domínio das informações, sendo capaz de explorar como esse conceito pode contribuir para melhorar a semântica em sistemas de recuperação da informação.

Borko (1968), ao discorrer acerca da Ciência da Informação, aponta a importância dos estudos interdisciplinares com a Ciência da Computação, além dos estudos de processamento e de técnicas aplicadas aos computadores. Embasado nessa ótica de Borko, este trabalho utiliza essa interdisciplinaridade entre a Ciência da Informação e a Ciência da Computação, na busca de compreender os processos e as teorias que são abarcados por essas duas ciências, tendo como prisma inicial os problemas inerentes à Ciência da Informação.

Vale destacar que a relação interdisciplinar entre Ciência da Computação e Ciência da Informação pode se vincular por meio da recuperação da informação e da Web Semântica, sendo um dos aspectos que guia o processo de desenvolvimento deste trabalho.

Dessa forma, o presente trabalho visa a trazer contribuições nas áreas de recuperação da informação e de Web Semântica, embasando-se nas técnicas da Inteligência Artificial e processamento de linguagem natural, bem como no uso de ontologias, para melhorar a contextualização das informações, aprimorando os sistemas de recuperação da informação no atendimento às necessidades informacionais dos usuários.

Repensar os modelos tradicionais de recuperação da informação, aproximando o usuário das Tecnologias da Informação e Comunicação e tornando as interfaces e o seu modo de uso cada vez mais natural, torna-se essencial para que a Ciência da Informação contribua em proporcionar uma recuperação mais próxima dos usuários, uma vez que serão utilizados os aspectos da linguagem humana para favorecer o encontro das informações que satisfaçam as necessidades informacionais durante o uso das tecnologias.

## 1.1 PROBLEMA

A distância existente entre a linguagem natural dos usuários e a linguagem computacional das Tecnologias da Informação e Comunicação cria uma dificuldade no uso de sistemas de recuperação da informação, devido à necessidade dos usuários se adequarem a expressões e sintaxes para realizar o processo da busca. Esse cenário, quando aliado ao grande volume informacional disponível na Web, torna os processos de recuperação da informação menos precisos e menos adequados aos usuários.

Ao aprofundar ainda mais a lacuna existente entre mecanismos de recuperação da informação e as necessidades informacionais dos usuários, evidencia-se uma dificuldade por parte desses mecanismos em compreender o sentido que as informações disponibilizadas na Web e em outras fontes informacionais possuem.

Sendo assim, surge a necessidade de aproximar a linguagem computacional da linguagem humana, para que o processo de recuperação da informação seja mais claro e eficiente para os usuários.

Nesse sentido, esta pesquisa se vincula às dificuldades durante o processo de recuperação da informação, no que tange, especialmente, à contextualização das informações pelos mecanismos computacionais, bem como, ao desafio existente de realizar o processo de busca e recuperação utilizando linguagem natural. Nesse contexto, aponta-se que os tradicionais

sistemas de recuperação da informação que realizam o processo de busca por meio da definição de palavras podem induzir os mecanismos computacionais a realizar a recuperação utilizando termos com expressividade limitada, tornando o conjunto de informações recuperadas pouco significativo para os usuários.

Como questões secundárias a serem respondidas, têm-se:

- Como os conceitos e as ferramentas da Web Semântica, em especial as ontologias, podem contextualizar os dados oriundos das fontes informacionais utilizadas em um sistema de recuperação da informação?
- Como um sistema de recuperação da informação baseado no processamento de linguagem natural pode utilizar as ferramentas e os conceitos da Web Semântica, para atender com mais eficiência as necessidades informacionais dos usuários, partindo de buscas com o uso de linguagem natural?
- Como o uso de bases de dados armazenados em contexto semântico pode auxiliar no processo de recuperação da informação?

## 1.2 TESE

A partir da problemática traçada, esta pesquisa defende a tese de que a recuperação da informação, quando realizada com o apoio do uso de linguagem natural, assim como de Inteligência Artificial e com a compreensão das propriedades semânticas dos termos, é capaz de tornar o atendimento às necessidades informacionais dos usuários mais eficiente e mais natural, atingindo diretamente o modo como as pessoas recuperam informações e permitindo um avanço na Ciência da Informação no que tange ao uso da Inteligência Artificial, possibilitando, também, repensar o modo como se compreende a recuperação da informação.

## 1.3 HIPÓTESE

Este trabalho apresenta a seguinte hipótese: se houver a união entre os conceitos e as ferramentas da Web Semântica e as técnicas relacionadas à Inteligência Artificial e ao processamento de linguagem natural, poderá ocorrer uma apropriação por parte dos sistemas de recuperação da informação da compreensão das necessidades informacionais dos usuários.

Tal hipótese é concebida pois a Web Semântica visa  a aprimorar a compreensão das tecnologias computacionais perante os conteúdos criados pelas pessoas, enquanto o

processamento de linguagem natural tem como meta processar e entender computacionalmente os termos de linguagem natural.

Adicionalmente, aponta-se a Inteligência Artificial como um meio de tornar a compreensão dos termos e dos conceitos de forma mais clara, possibilitando que o processo de recuperação da informação ocorra embasado em elementos que utilizem informações semânticas que aprimorem o processo de busca. Além disso, a Inteligência Artificial pode contribuir na aproximação entre os instrumentos computacionais e a linguagem natural, uma vez que esse campo de estudos é capaz de fornecer informações importantes sobre conceitos e termos, a partir de um histórico e com o apoio do aprendizado de máquinas, para aperfeiçoar o funcionamento dos modelos.

Dessa forma, essa aproximação no âmbito da Ciência da Informação possibilita aprimorar a recuperação da informação, para que, assim, as necessidades informacionais dos usuários possam ser mais bem atendidas.

## 1.4 PROPOSIÇÃO

Este trabalho se propõe a criar um modelo conceitual de recuperação da informação, centrado no uso de ferramentas da Web Semântica, do processamento de linguagem natural e da Inteligência Artificial, bem como na contextualização da informação recuperada. O modelo proposto estará embasado nos conceitos da Web Semântica, da Inteligência Artificial e da recuperação da informação.

## 1.5 OBJETIVOS

A explanação dos objetivos foi dividida em duas partes, objetivos gerais e objetivos específicos, descritos na sequência.

### 1.5.1 Objetivo geral

A pesquisa apresenta como objetivo geral a proposição de um modelo conceitual de recuperação da informação, a partir da aproximação da linguagem computacional com a linguagem natural, utilizando os princípios da representação da informação, para que o significado e o contexto dos dados estejam explícitos para o processo da busca. Para tal,

aproximam-se e relacionam-se os processos de Inteligência Artificial, processamento de linguagem natural e as ferramentas da Web Semântica.

### 1.5.2 Objetivos específicos

Como objetivos específicos busca-se:

- Conceituar e relacionar os conceitos de recuperação da informação, Web Semântica, Inteligência Artificial e processamento de linguagem natural.
- Definir tecnologias da Web Semântica e da Inteligência Artificial para a conceituação do modelo.
- Analisar como a Inteligência Artificial pode apoiar o processo de recuperação da informação.
- Definir como as ontologias poderiam ser utilizadas para a contextualização dos dados nas fases da recuperação da informação.
- Integrar o *Linked Data* como fonte de informação para auxiliar o processo de expansão da busca.
- Validar o modelo proposto por meio de uma prova de conceito que valide o modelo proposto.

## 1.6 JUSTIFICATIVA

A explosão informacional vivida no final do século XX e início do século XXI aumentou a necessidade de pesquisas que propusessem novas formas de recuperar as informações com mais precisão. A exigência de serem encontradas soluções que atuem dentro desse grande conjunto de dados, visando a atender com mais eficiência as necessidades informacionais dos usuários, é cada vez mais notória.

Dentro dessa seara, a Ciência da Informação tem um papel central, pois é necessário aprimorar os processos de recuperação da informação, utilizando as teorias informacionais e considerando os usuários, bem como se embasando em questões relativas à organização e à representação da informação.

Estudos buscando aperfeiçoar os processos de recuperação da informação no contexto da Web estão pesquisando novos meios de tornar mais precisos os seus procedimentos. Nesse

sentido, um dos caminhos que a disciplina de recuperação da informação está adotando é o de estudos que tratam do uso de ontologias em seus processos.

Essa tendência apresenta uma forte influência de pesquisas focadas nos conceitos, nas tecnologias e nas ferramentas da Web Semântica, pela emergência de que haja uma contextualização das informações, para que os sistemas computacionais sejam capazes de entender o significado dos dados disponíveis digitalmente.

Além disso, métodos de recuperação da informação, tal como o *Question Answering*, podem promover uma solução diferente para que os usuários possam sanar suas necessidades informacionais em ambientes digitais. No que se refere ao *Question Answering*, o *Linked Data* pode ser utilizado como uma fonte informacional, que reúne aspectos capazes de tornar a recuperação da informação mais eficiente, uma vez que são utilizados dados contextualizados e estruturados.

Dessa forma, esta pesquisa se justifica ao propor um novo meio de realizar a recuperação da informação, sob a perspectiva da Ciência da Informação, utilizando conceitos, tecnologias e ferramentas da Web Semântica, *Linked Data*, Inteligência Artificial e processamento de linguagem natural, visando a tornar o processo de recuperação natural e com um nível semântico elevado. A partir desse contexto, tem-se uma melhora no acesso à informação e na interação do humano com a máquina, uma vez que os usuários conseguiriam recuperar documentos e as informações de modo mais simples e natural.

Adicionalmente, ressalta-se que ao reunir elementos semânticos, especialmente as ontologias, com a recuperação da informação, baseada no *Question Answering*, juntamente com o *Linked Data*, como fonte informacional, obtém-se um diferente paradigma de recuperação da informação, que por meio das ontologias é capaz de fornecer a contextualização das informações, pela Inteligência Artificial pode oferecer respostas adequadas às questões realizadas pelos usuários e que encontra no *Linked Data* uma fonte de informação estruturada, construída com o conhecimento de milhares de organizações.

Vale destacar que as tecnologias estão evoluindo para uma aproximação da linguagem natural, pois os usuários se encontram cada vez mais imersos em seus dispositivos. Dessa forma, meios de recuperação da informação que utilizam o processamento de linguagem natural permitem uma aproximação do usuário com as TIC, inclusive possibilitando que as necessidades informacionais dos usuários sejam mais bem atendidas.

Por mais que estudos com sistemas de processamento de linguagem natural sejam feitos há várias décadas, uma apropriação desse campo de estudos por pesquisas de Web Semântica na Ciência da Informação permitirá que as pesquisas de recuperação da informação avancem

significativamente na direção de compreender, com mais propriedade, as necessidades informacionais dos usuários.

## 1.7 METODOLOGIA

Para o desenvolvimento desta pesquisa será utilizado o método quadripolar, proposto por Bruyne, Herman e Schoutheete em 1974. Essa proposta divide em quatro polos, os métodos de uma pesquisa: polo epistemológico, polo teórico, polo técnico e polo morfológico.

Os autores afirmam que tais polos "[...] não configuram momentos separados da pesquisa, mas aspectos particulares de uma mesma realidade de produção de discursos e de práticas científicos. Toda pesquisa engaja, explícita ou implicitamente, estas diversas instâncias [...]" (BRUYNE; HERMAN; SCHOUTHEETE, 1991, p. 35).

O método quadripolar foi assim proposto no âmbito das Ciências Sociais, visando a romper com uma linearidade na pesquisa, introduzindo um modelo topológico e não cronológico, segundo os autores supracitados. Posteriormente, em 2002, Silva e Ribeiro (2002) inseriram esse método como um dispositivo metodológico para a Ciência da Informação.

Silva (2017) afirma que a Ciência da Informação necessitava de um aperfeiçoamento em seus dispositivos metodológicos e a inserção do método quadripolar, como um instrumento para tal aperfeiçoamento, pode ser considerada passo estratégico e fundamental. O autor afirma que na Ciência da Informação, o método quadripolar dá "[...] especial ênfase ao polo epistemológico, o qual é apontado como a instância em que vigora o paradigma dominante até ser lentamente substituído por um outro [...]; e ao polo teórico [...]" (SILVA, 2014, p. 33)

Mais recentemente, alguns autores inseriram essa metodologia nas pesquisas focadas em informação e tecnologias no contexto da Ciência da Informação. Nesse cenário, Oliveira e Vidotti (2015, p. 37) afirmam que:

> O Método Quadripolar possui caráter dinâmico e flexível, qualidades que o tornam pertinente para ser usado em pesquisas que assumem a complexidade e a multidimensionalidade como elementos norteadores, a nosso ver, nas pesquisas que tratam informação e tecnologia de forma dialógica.

Esses autores, ao inserirem o método quadripolar no contexto de informação e tecnologia, auxiliam pesquisas científicas dessa temática a terem metodologias mais consistentes, que possam se adaptar a cada necessidade e aos distintos objetos estudados. Tal questão se mostra importante para esse domínio, pois insere uma metodologia consolidada nas áreas das Ciências Humanas e Sociais, para esse campo interdisciplinar da Ciência da Informação.

Silva (2006, p. 154, grifo nosso) apresenta os quatro polos:

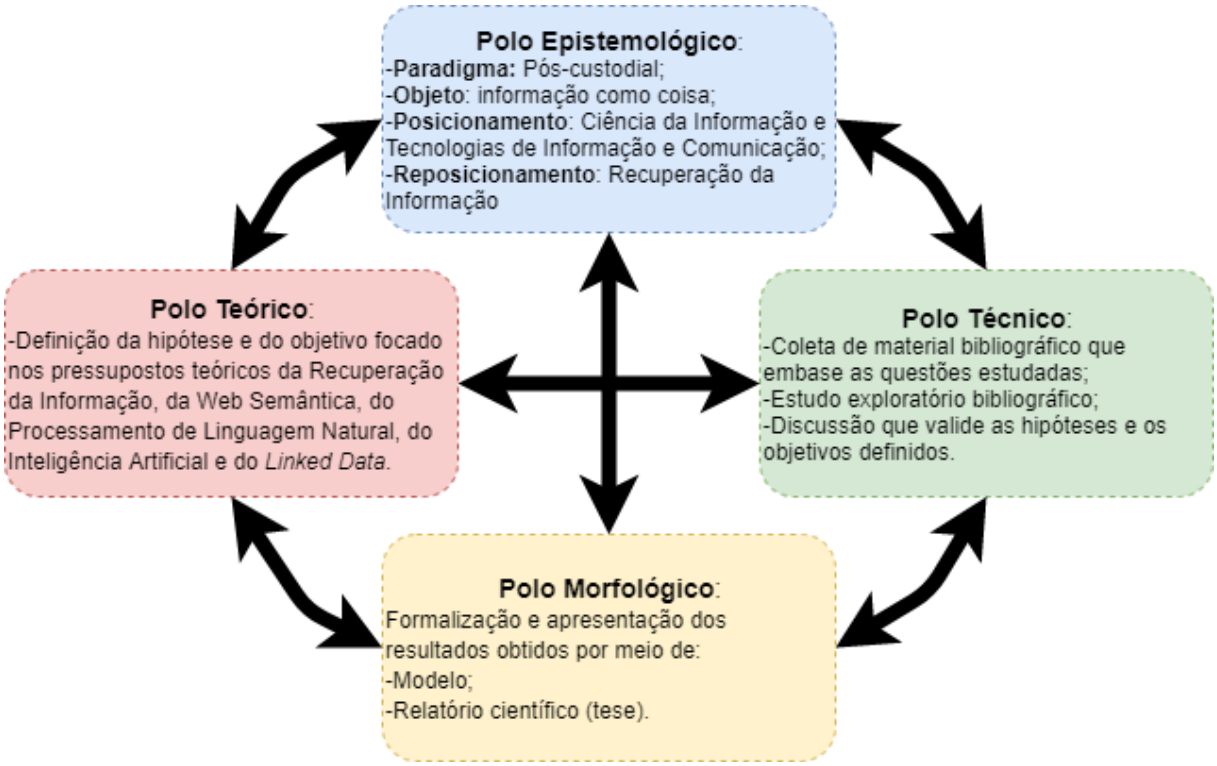
> No **pólo epistemológico**, opera-se a permanente construção do objecto científico e a definição dos limites da problemática de investigação, dando-se uma constante reformulação dos parâmetros discursivos, dos paradigmas e dos critérios de cientificidade que orientam todo o processo de investigação; no **pólo teórico**, centra-se a racionalidade do sujeito que conhece e aborda o objeto, bem como a postulação de leis, a formulação de hipóteses, teorias e conceitos operatórios e consequente confirmação ou informação do "contexto teórico" elaborado; no **pólo técnico**, consuma-se, por via instrumental, o contacto com a realidade objectivada, aferindo-se a capacidade de validação do dispositivo metodológico, sendo aqui que se desenvolvem operações cruciais como a observação de casos e de variáveis e a avaliação retrospectiva e prospectiva, sempre tendo em vista a confirmação ou refutação das leis postuladas, das teorias elaboradas e dos conceitos operatórios formulados; no **pólo morfológico**, formalizam-se os resultados da investigação levada a cabo, através da representação do objecto em estudo e da exposição de todo o processo de pesquisa e análise que permitiu a construção científica em torno dele.

Os quatro polos conduzem a uma clareza dos processos a serem realizados, ao mesmo tempo que rompem com o paradigma linear de uma pesquisa científica (VECHIATO, 2013). Silva (2006, p. 29) relata ainda que: "Nesta dinâmica quadripolar de investigação, assume particular relevância o polo teórico, uma vez que ele respalda a componente técnica e instrumental e dá sentido à explanação de resultados que consubstancia o polo morfológico."

Nesse contexto, a aplicação desse método no âmbito deste trabalho permite que uma metodologia consistente embase os processos da pesquisa científica aqui realizados, tanto em seu caráter teórico, quanto aplicado,

Assim, a figura 1 apresenta a inserção das questões relativas à pesquisa dentro dos polos do método quadripolar.

**Figura 1** - Delimitação da pesquisa dentro do método quadripolar




Fonte: elaborada pelo autor.

No polo epistemológico deverão ser tratadas as interseções e relações nas questões epistemológicas relativas à Ciência da Informação, para que assim esta pesquisa possa se embasar cientificamente e delimitar com clareza a problemática e as hipóteses a serem fornecidas. Nesse sentido, o presente trabalho parte do paradigma pós-custodial, informacional e científico apontado por Silva (2006, p. 158), que o define como um paradigma: "[...] emergente porque está a surgir no dealbar, em curso na era da informação e nos meandros de uma conjuntura de transição bastante híbrida, complexa e sujeita a um ritmo de inovação tecnológica e científica quase vertiginoso (a sociedade da informação, em rede etc.).". A pesquisa se encontra nesse paradigma, pois a complexidade dos ambientes informacionais digitais, no que tange a favorecer uma recuperação da informação mais próxima da linguagem natural, contempla as questões elencadas pelo autor.

Complementarmente, o objeto da pesquisa é a informação como coisa, definido por Buckland (1991). Utiliza-se esse conceito, pois será tratado apenas da informação registrada, que está colocada em algum tipo de suporte e é tangível originalmente. Ainda no polo epistemológico, esta pesquisa está posicionada no âmbito das Tecnologias de Informação e Comunicação, ao tratar das questões relativas aos estudos de tecnologias dentro da Ciência da Informação, em especial nas pesquisas de informação e tecnologia. Busca-se também

reposicionar epistemologicamente a recuperação da informação na Ciência da Informação, ao trazer contributos que complementam e ampliam o prisma da Ciência da Informação sobre esse campo de estudos. Esse reposicionamento busca inserir a recuperação da informação sob uma nova perspectiva, em que os processos de recuperação são revisados a partir da adoção e da inserção da Inteligência Artificial e do processamento de linguagem natural, de modo a possibilitar uma nova inserção da recuperação da informação na Ciência da Informação.

No que tange ao polo teórico, as hipóteses e os objetivos desta pesquisa serão delineados a partir de uma investigação teórica acerca das temáticas do *Linked Data*, das ontologias, da recuperação da informação, da Inteligência Artificial, da Web Semântica e do *Question Answering*, para que a partir dessa abordagem possa ser desenvolvido este trabalho. A pesquisa teórica dessas temáticas visa a permitir identificar e aprofundar a compreensão das ferramentas da Web Semântica, em especial da linguagem OWL, visando a traçar uma relação entre essas tecnologias com técnicas de recuperação da informação, utilizando processamento de linguagem natural, fundamentalmente o *Question Answering*. Associado a isso, tem-se o estudo sobre os princípios do *Linked Data*, para que estes possam posteriormente ser integrados aos processos de recuperação da informação.

O polo técnico reúne as tarefas que possibilitaram a realização da pesquisa, que permitirá validar ou não as hipóteses propostas inicialmente. Nesse sentido, esta pesquisa apresenta-se como um estudo exploratório bibliográfico, pois busca, na literatura e em documentos, embasamento para a construção de um modelo conceitual. Vale destacar que serão investigadas com mais profundidade questões relativas às relações do *Linked Data*, às propriedades do OWL, às formas de inserir o *Linked Data* e as ontologias OWL dentro do *Question Answering*, propondo o modelo conceitual e sendo validado por uma prova de conceito.

Por fim, no polo morfológico são apresentados os resultados obtidos com a pesquisa, que são em suma: o modelo conceitual, contemplando as características e os princípios teóricos que irão relacionar ontologias, *Question Answering* e *Linked Data*; em um momento posterior foi realizada uma prova de conceito validando o modelo, que também é apresentada, assim como o relatório científico, no formato de uma tese que será defendida no contexto do doutoramento do autor, explicitando o percurso da pesquisa científica desenvolvida.

## 1.8 TRABALHOS RELACIONADOS

Há muitas pesquisas que propuseram modelos de QA nas mais diversas áreas do conhecimento. Para realizar essa seção buscaram-se trabalhos que contemplavam as temáticas de "*Question Answering*", "Ontologias", "*Ontologies*", "*Linked Data*", "Recuperação da informação" e/ou "*Information retrieval*". Buscaram-se trabalhos que estavam indexados na Web of Science, BRAPCI, Portal de Periódicos da Capes e Google Scholar, publicados a partir de 2007, dez anos antes do início da realização deste trabalho (2017).

Ressalta-se que o número de trabalhos que tratam de QA dentro de cenários de *Linked Data* é restrito, havendo principalmente textos dentro da área da Ciência da Computação, estudos criando QA no *Linked Data*. Destacam-se os estudos de Lopez et al. (2013) e Walter (2012), que apresentam um número considerável de citações e que propõem modelos e implementam protótipos de QA em distintos ambientes do *Linked Data*. Essas pesquisas têm como centro o uso dos dados RDF do *Linked Data* como fonte de informação, realizando uma transformação de uma pergunta em linguagem natural em triplas RDF, e utilizando o SPARQL para buscar respostas, em bases de dados estruturados do *Linked Data*.

Em contrapartida, há algumas pesquisas que trabalharam com ontologias no contexto do QA. Em destaque, considere-se Asiaee et al. (2015), que implementam um sistema de QA com o uso de ontologias no contexto de dados de parasitas e a pesquisa de Kumar e Zayaras (2015), que propõem a construção de uma ontologia dinâmica, que contempla o conhecimento obtido a partir das perguntas realizadas no sistema de QA. Essas três pesquisas evidenciam, sob a ótica da Ciência da Computação, como as ontologias podem ser utilizadas para agregar compreensão da semântica dos termos nos sistemas de QA, utilizando ontologias no processo de interpretação de uma pergunta.

No que se refere aos trabalhos tratando de processamento de linguagem natural, e que está relacionado ao trabalho apontam-se dois trabalhos de caráter mais aplicado, demonstrando a popularização e o uso do PLN em diversas áreas. Um primeiro trabalho é de Bernardo e Santanché (2017) que desenvolveu um *chatbot* no desenvolvimento de jogos de saúde. Os autores fizeram um jogo que treina alunos de medicina na tomada de decisões em alguns quadros. Para isso, além do desenvolvimento e do uso de técnicas de PLN, utilizou-se uma ferramenta de conversa instantânea que já vem instalada nos telefones, facilitando, assim, a disseminação da aplicação.

Um outro trabalho foi desenvolvido por Costa, Campelo e Campos (2018), na área da educação. Os autores realizaram a classificação automática de questões matemáticas, que estavam de alguma forma ligadas ao pensamento computacional. Esse processo acontecia de

forma automática devido ao uso de técnicas de processamento de linguagem natural, além de aprendizagem de máquinas, reafirmando o momento atual da área de PLN, como relatado na subseção sobre o histórico dessa área.

Um outro recente trabalho é de Abdi, Idris, Ahmad (2018), que criam e apresentam um sistema de *Question Answering*, utilizando ontologias para a realização de inferências semânticas, dentro do domínio da física. Esse trabalho se diferencia da presente tese pelo fato de as ontologias no trabalho serem utilizadas para, a partir dos termos de textos, auxiliar na realização de inferências no processamento da questão e, nesta tese, as ontologias irão participar de todos os processos, utilizando, para isso, a capacidade semântica do *Linked Data*. Assim, ao trabalhar com bases nas estruturas de *Linked Data*, esta tese busca utilizar com maior aprofundamento a capacidade das ontologias.

Em um trabalho com um número considerável de citações, Ferrandéz et al. (2009) criou um sistema de *Question Answering* com o uso de ontologias para o domínio do cinema. Semelhante aos demais trabalhos, a ontologia é utilizada como uma representação do domínio do qual o sistema trata. No entanto, destaca-se esse trabalho, pois, a ontologia auxilia na compreensão de como o usuário escreveu a sua pergunta, para melhorar a sua interpretação do que deseja ao realizar a busca. Também esse trabalho se diferencia desta tese, visto que as ontologias não estão no centro dos três processos do *Question Answering*, para aprimorar e aumentar o nível de semântica formal em todo o processo.

Mais recentemente, Xie et al. (2015) desenvolveram um sistema de *Question Answering* que, para aprimorar os resultados do sistema, utilizou ontologias. Os autores utilizaram o Jena para realizar inferências e melhorar a interpretação semântica dos termos que os usuários estavam utilizando no sistema. Novamente, as ontologias vêm para apoiar um processo do QA, não estando, porém, presente em todos os processos que compõem esse tipo de recuperação da informação.

Em uma outra perspectiva, relacionando mais o *Question Answering e o Linked Data*, Bouziane (2018) traz uma importante contribuição para o uso do QA em bases de dados ligadas. Esse trabalho visa a permitir o uso de uma base em árabe e, para isso, utiliza o sistema de QA para explorar a Web Semântica árabe. Apesar de utilizar uma ontologia no processo, esse trabalho não está focado em aprimorar o nível de semântica com a utilização de ontologias nos diversos processos que compõem o QA.

Por fim, o último trabalho destacado aqui é o de Cabaleiro, Peñas e Manandhar (2017), que propõem uma metodologia para avaliar como as relações semânticas podem ser utilizadas, extrapolando o sentido da palavra tomada isoladamente, ou da unidade lexical. Destaca-se esse

trabalho, pois os autores citam o *Linked Data* como uma forma de verificar isso, devido a capacidade semântica de suas bases. Diferentemente da presente tese, as ontologias não são citadas como parte do processo, mas há apenas uma avaliação semântica das ligações do *Linked Data*.

O quadro 1 apresenta uma síntese dos trabalhos relatados anteriormente. Para isso os trabalhos foram divididos em 4 tipos, que refletem o aspecto principal de cada publicação. Os tipos são: 1) QA e *Linked Data*: esses trabalhos têm um enfoque maior no uso de dados, seguindo formatos de *Linked Data* para apoiar sistemas de *Question Answering*; 2) Ontologia e QA: trata de como as ontologias podem auxiliar no tratamento da linguagem nos sistemas de *Question Answering*; 3) Assistente Virtual: trabalhos que exploram o conceito e criam assistentes virtuais, e que têm em sua base o processamento de linguagem natural como essência; e 4) *Linked Data*, QA e ontologia: trabalhos que relacionam o *Linked Data* e as ontologias no âmbito do *Question Answering*. Ao final apresenta-se a comparação com o presente trabalho.

**Quadro 1** – Comparação de trabalhos relacionados

| **Tipo** | **Trabalho** | **Adoção *Linked Data*** | **Uso de ferramentas da Web Semântica** | **Ontologias em todos os processos do Q.A.** | **PLN para compreen-são da semântica** |
|---|---|---|---|---|---|
| QA e *Linked Data* | Lopez et al. (2013) | Sim | Sim | Não | Parcial |
| QA e *Linked Data* | Walter (2012) | Sim | Sim | Não | Parcial |
| Ontologia e QA | Asiaee et al. (2015) | Não | Sim | Não | Parcial |
| Ontologia e QA | Kumar e Zayaras (2015) | Não | Sim | Não | Parcial |
| Assistente virtual | Bernardo e Santanché (2017) | Não | Não | Não | Sim |
| Assistente virtual | Costa, Campelo e Campos (2018) | Não | Não | Não | Sim |
| Ontologia e QA | Abdi, Idris, Ahmad (2018) | Não | Sim | Não | Parcial |
| Ontologia e QA | Ferrandéz et al. (2009) | Não | Sim | Não | Parcial |
| Ontologia e QA | Xie et al. (2015) | Não | Sim | Não | Parcial |

| **Tipo** | **Trabalho** | **Adoção *Linked Data*** | **Uso de ferramentas da Web Semântica** | **Ontologias em todos os processos do Q.A.** | **PLN para compreen-são da semântica** |
|---|---|---|---|---|---|
| *Linked Data*, QA e ontologia | Bouziane (2018) | Sim | Sim | Não | Parcial |
| QA e *Linked Data* | Cabaleiro, Peñas e Manandhar (2017), | Sim | Parcial | Não | Parcial |
| | Coneglian (2020) | Sim | Sim | Sim | Sim |

Fonte: elaborado pelo autor.

Todos os trabalhos trazem importantes contribuições para a área de QA e PLN, porém nenhuma das pesquisas aborda, sob a ótica da Ciência da Informação, como as ontologias podem ser utilizadas durante os três processos de QA (processamento da pergunta, do documento e da resposta), sendo, em função disso, limitado o uso de ontologias como um auxílio para interpretar termos desconhecidos. Nesse sentido, o presente trabalho busca inserir as ontologias como o centro de QA, sendo utilizadas para contextualizar os termos escolhidos pelos usuários, os termos recuperados do *Linked Data* e a construção das respostas.

O uso das ontologias OWL em todos os processos poderá transformar o processo semântico como um todo, pois trabalhará inclusive na identificação dos resultados que possuem significados semelhantes aos da busca feita pelo usuário. As principais diferenças deste trabalho, com relação aos outros apresentados, estão primeiramente no uso de *Linked Data* juntamente com as ontologias OWL e, principalmente, pela utilização das ontologias como um mecanismo que vai além de buscas por palavras sinônimas, utilizando toda a capacidade semântica que o OWL possui na contextualização da informação, na criação de relações e na possibilidade da realização de inferências em todos os processos do QA.

No âmbito da Ciência da Informação, os trabalhos mencionados não trazem a conceituação e o uso de recuperação e representação da informação, enquanto que este demonstra como a definição do modelo proposto, considerando os aspectos e os modelos da Ciência da Informação, pode contribuir para aprimorar o processo da recuperação da informação, com o auxílio de Inteligência Artificial e processamento de linguagem natural.

# 2 RECUPERAÇÃO DA INFORMAÇÃO

O desenvolvimento deste trabalho está embasado em um conceito-chave, que é a recuperação da informação. Nesse sentido, esta seção destaca e apresenta a área da recuperação da informação, além de mostrar como a representação da informação pode auxiliar nesse processo.

A recuperação da informação, por ser o tema da principal contribuição deste trabalho, apresenta-se como área que possibilita as interdisciplinaridades e, em especial, permite uma conexão entre a Ciência da Informação e a Ciência da Computação. Ademais, ao tratar de questões como a Inteligência Artificial e o processamento de linguagem natural, a recuperação da informação pode ser aprimorada e melhorada.

Dessa forma, esta seção apresenta os conceitos de recuperação da informação, o conceito de relevância, os principais modelos de recuperação da informação, a aplicação de novas tecnologias nos modelos de recuperação da informação e, por fim, a relação entre representação da informação e recuperação.

## 2.1 CONCEITO DE RECUPERAÇÃO DA INFORMAÇÃO

Os estudos de recuperação da informação apresentam uma grande influência e relação com a Ciência da Informação. Em um clássico texto da área de Ciência da Informação, Saracevic (1995, p. 3, tradução nossa) discorre sobre as interdisciplinaridades dessa área, apontando que a recuperação da informação “[...] não é somente uma atividade da ciência da informação, mas a mais importante delas e também aquela em que mais ocorrem as relações interdisciplinares.”

O termo recuperação da informação (*information retrieval*) foi definido por Mooers em 1951, indicando que esse campo de estudos se preocupa com questões que perpassam pela descrição da informação, a especificação de busca, bem como os sistemas e técnicas utilizados para localizar uma informação. (MOOERS, 1951).

A partir da definição desse termo, pesquisas em diversas áreas do conhecimento começaram a ser realizadas tratando de recuperação da informação, pois os processos relacionados a essa área são centrais para a maioria das atividades. Naturalmente, a Ciência da Informação assumiu um papel central nessas pesquisas, tendo a colaboração da Ciência da Computação na busca de oferecer sistemas de recuperação da informação mais eficientes, capazes de processar mais rapidamente as informações, além de aprimorar os processos visando a contemplar as necessidades informacionais dos usuários.

Em suma, a recuperação da informação busca utilizar-se de diversos artifícios para aperfeiçoar os processos que perpassam pela finalidade de localizar uma informação que satisfaça a um determinado usuário. (FERNEDA, 2012).

Em uma outra vertente, mais focada no âmbito da Ciência da Computação, Baeza-Yates e Ribeiro-Neto (1999, p. 1, tradução nossa) consideram que

> [...] a recuperação da informação está diretamente ligada à representação, armazenamento, organização e acesso aos itens de informação. Dizem que a representação e a organização dos itens de informação deveriam prover o uso e o fácil acesso à informação necessária ao usuário.

Essa definição apresenta alguns elementos que são chave para compreender o conceito de recuperação da informação e, em especial, as questões de representação, armazenamento, organização e acesso às informações. Esses elementos são essenciais para que a recuperação da informação ocorra, pois, todos esses elementos são parte do processo da recuperação.

Apesar da forte influência dos autores na área da Ciência da Computação, visualiza-se que os elementos elencados por eles estão intimamente ligados à área da Ciência da Informação, demonstrando que o processo de recuperação da informação necessita considerar uma série de fatores, ligados aos usuários, e ao modo como estes fazem o acesso às informações.

Nesse contexto, Fusco (2010) e Coneglian (2017) trazem importantes considerações acerca da recuperação da informação no contexto bibliográfico, além de apontar questões sobre o seu objetivo. Assim, os autores destacam que o propósito da recuperação da informação está em fornecer resultados que tenham real valor e com importância para os usuários, de modo que possa existir uma maior aderência na intersecção entre os itens bibliográficos e as necessidades informacionais dos usuários.

Dessa forma, ambas as referências apresentam uma perspectiva de como a recuperação da informação deve ser tratada e avaliada, sempre considerando se essa intersecção foi satisfatória ou não.

Aprofundando a relação entre as áreas de Ciência da Informação e Ciência da Computação no contexto da recuperação da informação, Santarem Segundo (2010, p. 26-27) aponta que:

> A Ciência da Informação e a Ciência da Computação aparecem como as ciências mais envolvidas com a busca pela melhoria da qualidade da informação recuperada. A Ciência da Informação apresenta uma visão mais metodológica e tem procurado estruturar os dados e criar métodos e modelos que proporcionem um melhor armazenamento da informação, assim como vem estudando métodos que agreguem semântica à informação, e consequentemente possam ser aplicadas no processo de recuperação. A Ciência da Computação tem procurado atuar na aplicação dos modelos

> citados, diretamente no desenvolvimento de técnicas computacionais, como algoritmos, que possam viabilizar as metodologias sugeridas e pesquisadas.

A visão apresentada pelo autor diferencia os contributos teóricos e metodológicos que as diferentes áreas dão para o campo da recuperação da informação. Nesse contexto, a área da Ciência da Computação tem apresentado mais soluções aplicadas, utilizando as técnicas computacionais e algoritmos, enquanto a Ciência da Informação realiza reflexões epistemológicas e metodológicas, criando modelos teóricos que visam a aumentar o nível semântico das soluções. Destaca-se que este é um papel complementar que ambas as áreas possuem, tendo uma profunda interdisciplinaridade nessa área, destacando as afirmações de Saracevic (1995), de que a recuperação da informação é o principal campo de interdisciplinaridade na área da Ciência da Informação.

Souza (2006, p. 172) destaca que a área da recuperação da informação traz ainda a necessidade de integração com outras áreas, relatando que: “Uma real integração demandaria estudos concomitantes em diferentes áreas do conhecimento [...] como a ciência da informação, a linguística, a ciência da computação, com a inteligência artificial; a psicologia cognitiva, a comunicação, a sociologia, a antropologia, entre outras.”

As áreas e campos de estudos apresentados pelo autor demonstram que a recuperação da informação necessita de integração com outros campos, trazendo uma tendência de interdisciplinaridade, que vem se consolidando nos últimos anos. Assim, apesar do pioneirismo da Ciência da Computação e da Ciência da Informação nesses estudos, há oportunidades de aprimorá-la e relacioná-la com outras ciências e áreas.

Outro importante aspecto está na perspectiva que a recuperação da informação pode ter. Nesse contexto, Baeza-Yates; Ribero-Neto, 1999, p. 1, tradução nossa) apontam que:

> Na visão centrada no computador, a RI [recuperação da informação] consiste principalmente na criação de índices eficientes, no processamento de consultas de usuários com alto desempenho e no desenvolvimento de algoritmos de classificação para melhorar os resultados. Na visão centrada no ser humano, a RI consiste principalmente em estudar o comportamento do usuário, em entender suas principais necessidades e em determinar como essa compreensão afeta a organização e a operação do sistema de recuperação.

A visão apresentada pelos autores demonstra que as duas perspectivas são complementares e devem ser tratadas nos estudos de recuperação da informação. Dessa forma, não basta ter um bom sistema, com ótimos índices, mas que não atende às necessidades dos usuários, ou que tenha uma usabilidade precária. Este trabalho busca atender as duas questões, uma vez que se busca aperfeiçoar tanto o processamento, ao inserir novas tecnologias no

processo, quanto aprimorar a relação entre sistema e usuário, ao oferecer um outro modo de permitir a consulta, com linguagem natural.

Destaca-se ainda que, ao discutir o campo da recuperação da informação, é necessário abordar os sistemas de recuperação da informação. Esses sistemas são definidos por Santarem Segundo (2010) como modelos complexos que são responsáveis por todo o sistema de representação, armazenamento, gestão e recuperação da informação.

Lancaster e Warner (1993) relatam que os sistemas de recuperação da informação realizam a interface entre as coleções e os recursos informacionais com os usuários no geral. Além disso, eles apontam que esses sistemas realizam as tarefas de aquisição e armazenamento de documentos, de organização e controle e de distribuição e disseminação aos usuários.

Esse tipo de sistema apresenta algumas características essenciais, que foram destacadas por Souza (2006, p. 163):

> • Representação das informações contidas nos documentos, usualmente através dos processos de indexação e descrição dos documentos; • Armazenamento e gestão física e/ou lógica desses documentos e de suas representações; • Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as necessidades de informação dos usuários. Para isso é necessário que haja uma interface na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações.

As três características destacadas passam por três elementos essenciais: a representação, o armazenamento e a recuperação da informação. Assim, reafirmando o que Santarem Segundo (2010) relata, demonstra-se que o sistema de recuperação da informação não está limitado ao processo de recuperar, sendo um processo mais holístico, que passa por diversos âmbitos dos dados e das informações.

Esses sistemas estão evoluindo significativamente nos últimos anos, trazendo conceitos da Web Semântica, processamento de linguagem natural e Inteligência Artificial, mas mantêm os elementos clássicos apresentados pelos autores. No contexto dessa evolução, Monteiro et al. (2017) relatam que os atuais SRI realizam processos mais complexos e sofisticados de coleta, indexação, análise e interfaces de busca, destacando que são utilizadas tecnologias diversas para realizar uma representação e descrição mais aprimorada. Os autores apontam ainda: “ontologias para relações de domínio; bases de conhecimento [...] que fornecem dados estruturados para os agentes inteligentes; programas e tecnologias de visualização e apresentação dessa informação em contextos [...] nas interfaces de busca.” (MONTEIRO et al., 2017, p. 165).

A questão apresentada pelos autores aponta que as evoluções das áreas de sistemas de recuperação da informação ocorrem nas diversas frentes, desde os próprios algoritmos de

recuperação, passando por melhorias na descrição e representação, chegando até as interfaces que passam a considerar o contexto da busca.

O processo de recuperação da informação está bastante vinculado ao conceito de relevância, essencial para definir as técnicas e modelos que podem ser utilizados para melhor satisfazer as necessidades informacionais dos usuários. Dessa forma, apresenta-se a seguir uma conceitualização da relevância no âmbito da recuperação da informação.

## 2.2 RELEVÂNCIA

A relevância é um dos principais elementos a serem considerados no momento de desenvolver um sistema de recuperação da informação, visando identificar como ele será tratado e os elementos considerados para definir como será ordenado e como será dado o acesso aos usuários.

Neste trabalho, o conceito de relevância permeia todo o processo da definição do processo de busca, visando a expandir a expressão ou pergunta realizada pelo usuário, para que o processo de recuperação da informação seja mais satisfatório.

Um dos principais autores a discutir o conceito de relevância do contexto no âmbito da Ciência da Informação foi Saracevic, em 1975, que, no texto "*Relevance: A review of and a framework for the thinking on the notion in information science*", discute as diferentes visões existentes sobre a relevância no campo da Ciência da Informação.

Saracevic (1975, p. 143, tradução nossa) resume como a relevância era tratada naquele momento, relatando que "Relevância é considerada como uma métrica da eficiência do contato entre uma fonte e um destino, em um processo de comunicação."

Nesse trecho, o autor demonstra a relação da relevância com a área da comunicação, evidenciando que um dos principais elementos  no que tange a relevância está na sua eficiência em atender às necessidades de informação de um destino (usuário). Ao pensar sob uma perspectiva da recuperação da informação, pode-se visualizar a necessidade de atender às necessidades informacionais de um usuário (destino), em busca nas fontes informacionais (fontes). Se esse processo tiver eficiência, é considerado que aconteceu de forma relevante.

O conceito de relevância foi discutido com bastante profundidade por Hjorland. Esse autor utiliza os conceitos da Ciência da Informação, aprofundando a compreensão acerca da relevância. Dessa forma, aponta que:

> [...] a relevância nunca é 'de um sistema', mas sempre 'humana' e, portanto, a dicotomia é errada. O determinar quais itens são relevantes em relação a uma determinada meta / tarefa, requer conhecimento do sujeito e é dependente de

> diferentes teorias / visões. Por conseguinte, os utilizadores dos sistemas de informação não são automaticamente competentes para julgar a pertinência. (HJORLAND, 2009, p. 231, tradução nossa)

A visão apresentada pelo autor considera que a relevância deve ter sempre como perspectiva o humano, pois determinar se algo foi ou é relevante, depende do que um determinado usuário espera, ou o tipo de informação necessária para ele. Assim, a visão destacada por Hjorland demonstra a complexidade da discussão acerca da relevância, e o quão difícil é a criação de sistemas que consigam atender às necessidades informacionais dos usuários, uma vez que cada usuário tem uma percepção, uma necessidade e um contexto.

Uma visão complementar à apresentada por Hjorland foi dada por Borlund em 2003. O autor, buscando partir de uma visão mais objetiva, que pudesse, de alguma forma, levar o conceito para os sistemas informacionais, dividiu a relevância em duas classes: "[...] (1) relevância objetiva ou baseada em sistema; e (2) relevância subjetiva ou humana (usuário)." (BORLUND, 2003, p. 914, tradução nossa).

O autor relata que diversas outras pesquisas, realizadas por especialistas e pesquisadores da área da Ciência da Informação, recuperação da informação, sumarizam essa questão da relevância nessas duas classes. Adicionalmente, Borlund (2003, p. 914, tradução nossa) afirma que:

> As duas principais classes de relevância são bastante diferentes por natureza e, por padrão, implicam diferentes graus de envolvimento intelectual. Cada uma das duas principais classes de relevância corresponde à compreensão da relevância empregada por cada uma das duas principais abordagens de pesquisa e avaliação de RI: as abordagens orientadas ao usuário e orientadas pelo sistema.

A visão do autor demonstra, por um lado, que a relevância orientada a sistemas tem uma perspectiva mais objetiva e estática, em que ficam claros os parâmetros e métricas para determinar aquilo que é relevante. Por outro lado, a relevância orientada ao usuário é mais complexa, uma vez que considera uma experiência mental individualizada, que é algo naturalmente subjetivo.

No contexto da recuperação da informação, Borlund (2003, p. 923, tradução nossa) retoma essa discussão trazendo que: "Abordamos a relevância com base na ideia do ponto de vista cognitivo, em que os julgamentos de relevância evoluem durante o processo de interação de RI."

O ponto apresentado pelo autor demonstra que a relevância, e o seu julgamento, no âmbito da recuperação da informação está em constante evolução, visto que há mudanças

quanto ao modo de ver o que é relevante, além de que outros modos de interação e visualização são aspectos que contribuem para compor esse cenário de relevância.

Mais recentemente, ao falar sobre mecanismos de buscas, alguns autores apontam alguns aspectos complementares que são levados em conta, ao considerar algo relevante. Nesse contexto, Monteiro et al. (2017, p. 173) relatam que:

> [...] o conceito de relevância confunde-se com a otimização semântica, isto é, com os significados e também com o menor esforço e maior efeito cognitivo do usuário, [...], a relevância é um conceito em devir, à espera de novas atualizações conceituais, cognitivas e técnicas. Em especial, nos mecanismos de busca, o conceito está fortemente relacionado às tecnologias semânticas do ciberespaço, à personalização e à contextualização da busca.

Os autores destacam alguns importantes elementos que estão sendo fortemente considerados no âmbito dos mecanismos de buscas. Em especial, as tecnologias semânticas, a personalização e a contextualização passam a ser elementos centrais para a definição da relevância no âmbito da Web. Nesse sentido, há uma forte influência das ferramentas da Web Semântica e da própria inserção do processamento de linguagem natural apoiando o processo de busca.

Vale destacar um outro importante elemento colocado por Monteiro et al. (2017) que está na atualização dela, do ponto de vista conceitual, cognitivo e técnico, dando-lhe um caráter um caráter objetivo, pois, a relevância depende do cenário atual e das tecnologias que se apresentam nestes dias. Assim, a relevância está em uma constante evolução, o que não indica que um sistema de recuperação da informação construído há algumas décadas não apresentava relevância, e algo feito nos dias atuais apresenta resultados relevantes; deve-se considerar, no entanto, que houve uma mudança quanto às tecnologias e quanto ao conceito de relevância.

Em uma perspectiva complementar às apresentadas, Silva, Santos e Ferneda (2013, p. 37) relatam que se "[...] torna difícil criar estruturas artificiais capazes de garantir que os resultados de uma busca sejam relevantes ao seu usuário. Resume-se, basicamente, em mostrar os resultados possivelmente mais relevantes em forma de ranque (ranking), do mais [...] ao menos relevante."

Os autores demonstram que a relevância, na prática, é refletida no modo como os resultados são ordenados ao serem apresentados para os usuários. Destaca-se que a relevância é mais complexa e leva em consideração uma série de outros fatores, mas o modo como isso é apresentado nas interfaces para os usuários, se dá nesse *ranking*. Vale ressaltar que há outros aspectos, como aqueles discutidos por Monteiro et al. (2017), que estão sendo inseridos nos

mecanismos de busca e recuperação da informação, que estão oferecendo outras possibilidades para esses tipos de sistemas de informação.

Esse conceito de relevância está, pois, diretamente relacionado aos modelos de recuperação da informação que irão impactar significativamente no atendimento ou não às necessidades informacionais dos usuários.

## 2.3 MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

Há diversos modelos de recuperação da informação que atendem a diferentes necessidades e foram desenvolvidos em várias épocas, desde o início desse campo de estudo. Destacam-se os modelos clássicos que influenciaram significativamente os modelos atuais e são utilizados em alguns sistemas até hoje.

Ferneda (2003, p. 18-19) afirma que "A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que o mesmo utiliza. Um modelo, por sua vez, influencia diretamente no modo de operação do sistema."

Há basicamente três modelos clássicos que serão apresentados a seguir: modelo booleano, modelo vetorial e modelo probabilístico. Tais modelos foram desenvolvidos nas décadas de 1960 e 1970 e foram aprimorados na década de 1980, sendo fundamentais para os estudos e as práticas de recuperação da informação.

O primeiro modelo, o booleano, é baseado na álgebra de boole, utilizando os operadores "AND" (E), "OR" (OU) e "NOT" (NOT) para realizar o processo de recuperação. Em suma, são utilizados esses operadores para encontrar dentro do corpus de documentos, aqueles que contêm determinadas palavras e não contêm outras. Caso o usuário deseje encontrar documentos que contenham duas ou mais palavras, utiliza-se o operador AND, e caso deseje encontrar um documento que contenha uma ou outra palavra, utiliza-se o operador OR.

O princípio desse modelo está na divisão do corpus de documentos em dois subconjuntos, o primeiro sendo aquele que atende à expressão de busca realizada pelo usuário, e o segundo aquele que não atende a essa expressão. Vale destacar que não há um critério de relevância nesse modelo, pois a única classificação realizada são esses dois subconjuntos, dificultando o atendimento, de forma satisfatória, das necessidades informacionais dos usuários.

Nesse contexto, Santarem Segundo (2010, p. 32) aponta que: "Não é possível, através do modelo booleano, apresentar resultados parciais, a estrutura binária de funcionamento sempre apresenta resultados exatos, baseados nas comparações binárias de 1 ou 0." O

argumento apresentado pelo autor reafirma a questão trazida, do modelo apenas considerar se algo atendeu ou não à expressão buscada, não havendo um meio termo e não podendo, assim, falar que uma resposta pode ser mais ou menos relevante. Sendo assim, pode-se afirmar com Ferneda (2003, p. 26): “Presume-se que todos os documentos recuperados são de igual utilidade para o usuário. Não há nenhum mecanismo pelo qual os documentos possam ser ordenados.”

A limitação do modelo booleano levou à necessidade de reflexões e da criação de outros modelos, capazes de aprofundar a compreensão do corpus de documentos, permitindo que mais expressões fossem realizadas. No entanto, há ainda diversos mecanismos que permitem a aplicação dos operadores e das técnicas oriundas desse modelo, juntamente com outras técnicas, modelos e algoritmos.

O segundo modelo clássico é o modelo vetorial, que tem como princípio o cálculo da similaridade entre um documento e uma expressão realizada por um usuário. Esse modelo foi proposto no ano de 1968, por Gerald Santon, principalmente pelas deficiências existentes no modelo booleano. (SANTAREM SEGUNDO, 2010).

Santarem Segundo (2010, p. 33) relata que a similaridade é representada por meio de:

> [...] um vetor numérico, onde cada elemento do vetor representa um termo de consulta e a este é atribuído um peso que indica tamanho e direção do vetor de representação. São esses pesos que possibilitam a proximidade de consulta e o cálculo da similaridade parcial entre os termos da consulta e os documentos, possibilitando que os resultados sejam apresentados de maneira classificada, de acordo com o grau de similaridade entre o termo na expressão de busca e o documento recuperado.

Em suma, o processo de definição da similaridade acontece por meio de uma fórmula, que utiliza o ângulo do vetor para definir qual a similaridade de um termo de busca e dos documentos que compõem o corpus. Utilizam-se como variáveis desse processo de definição da similaridade a quantidade total de documentos, a quantidade de ocorrências de um termo no corpus de documentos e a quantidade desse mesmo termo nos documentos de forma individual.

Em uma perspectiva de relevância, Souza (2006, p. 166-167) afirma que: “No modelo, que é não binário, pode-se calcular um grau de similaridade a ser satisfeito pelos documentos para serem considerados relevantes (ex: que as palavras apareçam ao menos duas vezes, etc.) e determinar o grau de similaridade, com vistas a construir um ranking.”

A partir do que foi relatado, pode-se considerar que o modelo vetorial possibilita a análise de outras variáveis, que permitem a inserção de relevância e a criação de ranking, o que não é possível com o modelo booleano. Ferneda (2003, p. 28) reafirma esse ponto, ao explicitar que: “Como resultado, obtém-se um conjunto de documentos ordenados pelo grau de similaridade de cada documento em relação à expressão de busca.”

O terceiro modelo clássico é o modelo probabilístico, que utiliza como princípio a probabilidade dos documentos atenderem às necessidades informacionais dos usuários. Tal modelo, aplicado à recuperação da informação, foi proposto por Maron e Kuhns, em 1960, realizando a classificação de documentos conforme a probabilidade de estarem relacionados aos termos de buscas dos usuários.

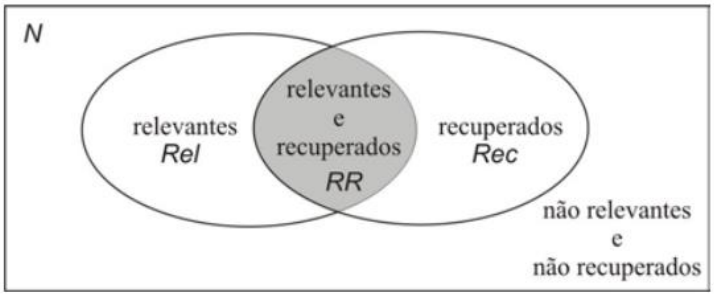
Coneglian (2017, p. 67) afirma que:

> Nesse modelo, a busca inicial apresenta um conjunto de documentos, e o usuário seleciona aqueles que ele considera relevantes, sendo armazenado esse feedback do usuário. Após sucessivas iterações, o sistema irá calcular quais são os documentos que têm a maior probabilidade de atender àquela nova busca, apresentando resultados baseados nos feedbacks recebidos.

Verifica-se que há um processo iterativo, que vai aprimorando os resultados, e com o passar do tempo os resultados alcançados são cada vez melhores, pois, a probabilidade de atender o que o usuário necessita aumenta, conforme os feedbacks vão aperfeiçoando a qualidade do processo.

A figura 2 aponta o conceito de como o corpus de documentos é dividido nesse modelo: há os documentos que são os não relevantes e não recuperados, os que são relevantes, os recuperados e, na interseção desses dois, os relevantes e recuperados.

**Figura 2** - Divisão do corpus de documentos no modelo probabilístico




Fonte: Ferneda (2003, p. 39).

Ferneda (2003), ao discutir a questão dessa divisão, considera que todo sistema de recuperação deve ter o maior número possível de documentos relevantes recuperados. O modelo probabilístico utiliza os conceitos probabilísticos e de feedback que os usuários dão para aumentar essa interseção; assim, em uma primeira iteração, são recuperados poucos documentos relevantes, mas à medida que vão ocorrendo as iterações, com os feedbacks fornecidos, a quantidade de documentos relevantes recuperados tende a aumentar. Ferneda

(2003, p. 39) relata ainda que: “Com os resultados obtidos após a execução da primeira busca é possível gradativamente melhorar os resultados através de interações com o usuário.”

Complementarmente, Santarem Segundo (2010, p. 35) afirma que:

> O modelo probabilístico caracteriza-se, principalmente, por apresentar um bom desempenho quando aplicado, visto que as estimativas de probabilidade já apresentam resultados de classificação, que podem ser utilizados para apresentação dos resultados; entretanto, é notável que o fato de não explorar a frequência dos termos é visto como ponto negativo do modelo.

A afirmação do autor evidencia o fato de que, embora bons resultados sejam alcançados com esse modelo, há problemas na sua aceitação; em especial, devido a não utilizar e explorar a quantidade de vezes que os termos aparecem nos documentos, como o modelo vetorial faz.

Em suma, os três modelos podem ser complementares, como podem, também, ser utilizados juntos, para obter melhores resultados. No geral, o modelo vetorial inspira a maioria dos modelos atuais, mas são também empregados conceitos do modelo booleano e elementos do modelo probabilístico, como o próprio feedback, para aumentar a probabilidade de satisfazer às necessidades dos usuários.

Há outros modelos de recuperação da informação, consequência da evolução das tecnologias e da área em si. Esses modelos utilizam alguns dos princípios dos modelos clássicos, inserindo outros elementos. Destaca-se o uso da lógica de fuzzy, o uso de redes neurais e algoritmos genéticos, e a aplicação de sistemas especialistas que são desenvolvidos para atenderem a cenários específicos, que será melhor explanado na subseção a seguir.

Uma síntese dos modelos de recuperação da informação e de sua evolução foi apresentado por Roa-Martinez (2019). O quadro 2 apresenta tal síntese.

**Quadro 2** – Síntese dos modelos de recuperação da informação:

| **MODELO** | **DESCRIÇÃO** |
|---|---|
| Modelo booleano (1950) | Baseado em álgebra booleana, a consulta e os recursos são considerados como um conjunto de termos. Um documento é relevante para uma consulta se contiver os termos de consulta. O modelo booleano é um modelo de correspondência exato, em que um recurso é relevante ou não para uma consulta dada. |
| Modelo vetorial (1970), proposto por Gerard Salton | Baseado na teoria do espaço vetorial, a consulta e os documentos são considerados vetores de termos. Um documento é relevante para uma consulta de acordo com o seu produto escalar (cosseno, coeficiente de dados, etc.). |
| Modelo probabilístico (1976), proposto por Robertson e Karen Spärck-Jones | Baseado em pressupostos probabilísticos, a consulta e os documentos são considerados como um conjunto de eventos. A relevância de um recurso para uma consulta depende da probabilidade condicional do documento |
| Modelo booleano estendido (1983), proposto por Gerard | Introduz a ponderação de termos na recuperação booleana. Este modelo usa representação vetorial e cálculo de distância entre |

| MODELO | DESCRIÇÃO |
|---|---|
| Salton, Edward A. Fox e Harry Wu | vetores para determinar a relevância de um documento versus uma consulta booleana |
| Conjuntos Fuzzy (1984), proposto por Lofti A. Zadeh | Introduz uma semelhança gradual entre os documentos e as consultas em modelos baseados em teorias de conjuntos. |
| De rede neural (1989), proposto por David E. Rumelhart e James McClelland | Representa dependências entre o termo e os aspectos dinâmicos da representação dos recursos usando a teoria neural. A consulta é o estímulo inicial da rede neural e a resposta da rede é o conjunto de documentos que são ativados pelos estímulos iniciais, porém, não há uma ativação usada como critério de relevância. |
| Modelo inferencial (1992) | Representa interdependências e dependências de termos de documentos. A relevância de um documento para uma consulta dada corresponde ao grau em que o documento satisfaz a necessidade do usuário. |
| Modelo de indexação semântica latente (1990), apresentado por Deerwester, Dumais, Furnas, Landauer e Harshman | Baseado na decomposição de valor singular que transforma o espaço vetorial do documento inicial em outro espaço vetorial de documentos semelhantes que estão mais próximos um do outro. |
| Modelo estrutural da linguagem (2006) | Incorpora as informações de estrutura do corpus usando agrupamentos (clusters) de documentos similares. |
| Modelo de grafos (2009) | Introduz a estrutura de grafo para determinar semelhança entre uma consulta e um documento. |

Fonte: Roa-Martinez (2019, p.45).

A evolução apresentada pela autora demonstra como os diversos modelos foram evoluindo e se complementando. Destaca-se que a aplicação de Inteligência Artificial em processo de recuperação da informação não é nova; considere-se, nesse sentido, os conjuntos de Fuzzy, que possuem uma relação com a Inteligência Artificial, assim como o das redes neurais em 1989.

Todo esse processo e novos modelos que foram surgindo evidencia que a área de recuperação da informação tem evoluído com novas tecnologias e conceitos que têm expandido o modo como os usuários buscam documentos e outras informações. Considere-se, nesse sentido, que, com a aplicação da Inteligência Artificial, a área de recuperação da informação está evoluindo com mais velocidade, e tendo melhores resultados. A seguir, apresenta-se com mais detalhes o impacto da Inteligência Artificial no processo de recuperação da informação.

## 2.4 INTELIGÊNCIA ARTIFICIAL NA RECUPERAÇÃO DA INFORMAÇÃO

O processo de recuperação da informação foi sendo alterado com influência das novas tecnologias computacionais. Destacam-se algumas das principais tecnologias que, nos últimos anos, estão influenciando significativamente o processo de recuperação da informação; em especial, devido a evolução das capacidades de processamento e armazenamento, que possibilitaram a aplicação de distintas técnicas de Inteligência Artificial.

A partir disso, tendo como enfoque abordagens da Inteligência Artificial, o quadro 3 apresenta algumas das novas técnicas utilizadas para a recuperação da informação.

**Quadro 3** – Aplicação da Inteligência Artificial na recuperação da informação

| **Técnica de Inteligência Artificial** | **Aplicação na recuperação da informação** | **Exemplos de aplicação** |
|---|---|---|
| Redes neurais | Rede neural é uma das mais tradicionais técnicas da Inteligência Artificial, simulando o funcionamento de um neurônio humano. Em suma, uma rede neural visa a identificar padrões, aprendendo de acordo com os dados que serão utilizados para o treinamento e a execução dos processos. Na recuperação da informação, as redes neurais recebem as expressões de busca visando a identificar quais são os melhores documentos para satisfazer às necessidades informacionais dos usuários. Um trabalho apresentado com a aplicação de redes neurais em recuperação da informação demonstrou como as redes neurais podem auxiliar nos processos de agrupamento e com aprendizado de máquinas. | Capuano (2009) |
| Algoritmos genéticos | Os algoritmos genéticos funcionam baseado nas teorias da biologia evolutiva. Na prática, os algoritmos genéticos irão evoluindo as populações, cruzando, mudando e eliminando os indivíduos, para que a nova geração seja mais adequada. Assim, as melhores opções são aprimoradas e os piores são eliminados, sendo assim, uma Inteligência Artificial que irá aperfeiçoar as opções de processamento. Na recuperação da informação, os algoritmos genéticos podem ser utilizados para definição de parâmetros e pesos para que o processo de recuperação seja melhor ajustado e adequado. Um exemplo de aplicação está na recuperação de imagens, em que os parâmetros e pesos para realizar buscas de imagens em redes sociais teve um bom resultado para o ajuste do processo de busca, além de possibilitar um constante ajuste. | Silva e Calumby (2018) |
| *Lógica Fuzzy* | A lógica Fuzzy é uma lógica em que os valores, ao contrário da lógica binária (apenas 0 e 1), podem assumir quaisquer valores reais entre 0 e 1. Assim, tem-se valores que podem estar mais próximos ou menos próximos de ser verdadeiros, o que está vinculado a uma verdade parcial. Esse processo, quando aplicado à recuperação da informação, permite que a lógica fuzzy seja adequada à subjetividade e imprecisão desse processo. Uma aplicação está no uso da lógica Fuzzy para as diversas partes de um documento estruturado, como um artigo científico, auxiliando assim, a definição da importância de cada parte do documento para a recuperação. | Ferneda e Dias (2013) |
| *Deep Learning* | A abordagem com *Deep Learning* utiliza pouco pré-processamento dos dados, visando ao aplicar, explorar a maior quantidade de camadas, e um nível profundo de análise dos dados. Na recuperação da informação, a utilização de *deep learning* para recuperação da informação como músicas e imagens tem-se mostrado bastante | Guimarães (2018) |

| Técnica de Inteligência Artificial | Aplicação na recuperação da informação | Exemplos de aplicação |
|---|---|---|
| | vantajosa. Isso ocorre, pois tanto as imagens quanto os áudios contêm grandes quantidades de informações, favorecendo com que o *deep learning*, por ser um aprendizado de máquinas que tem um grafo de análise mais profundo, seja aderente à recuperação da informação, que exige comparação com grandes quantidades de dados.<br>Em comparação com técnicas tradicionais de recuperação, uma aplicação para recuperação de músicas obteve um resultado mais satisfatório com o uso de *deep learning*. | |
| Processamento de linguagem natural | Uma das áreas da Inteligência Artificial que está afetando a área de recuperação da informação é o processamento de linguagem natural, que está modificando o modo como as pessoas interagem com os dispositivos computacionais, de modo que os indivíduos possam ter uma comunicação mais natural com os dispositivos. No âmbito da recuperação da informação, o uso de *Question Answering* favorece que a recuperação da informação ocorra com mais naturalidade, permitindo que o usuário faça uma busca fazendo uma pergunta à máquina, e a resposta seja dada em forma de uma resposta. Para realizar essa transformação, realiza-se busca em uma série de documentos, com uma exploração, classificação e filtro de parágrafos, para apresentar ao usuário a melhor resposta a uma pergunta. | Allam e Haggag (2012) |

Fonte: elaborado pelo autor.

O quadro 3 explicita as principais relações da Inteligência Artificial com a área de recuperação da informação. É possível verificar que a área de recuperação da informação está se adequando e se aprimorando com essas distintas técnicas, desde algumas mais antigas como as redes neurais e algoritmos genéticos, que ainda são bastante utilizados, até *deep learning* e processamento de linguagem natural, com o *Question Answering*, que tem modificado os processos da recuperação da informação para o uso maior de dados e de linguagem natural.

O quadro demonstra ainda que a área de recuperação da informação permite um aprofundamento com as técnicas de Inteligência Artificial, indo ao encontro da proposta do presente trabalho.

## 2.5 REPRESENTAÇÃO DA INFORMAÇÃO NA RECUPERAÇÃO DA INFORMAÇÃO

Um importante aspecto da recuperação da informação está na representação que os documentos possuem. O uso de metadados e outros elementos descritivos são essenciais para que os documentos possam ser recuperados com mais exatidão e eficiência, tendo em vista a

diversidade de elementos que podem ser utilizados para representar um determinado documento.

Em uma tradicional definição da representação da informação, Novelino (1996, p. 38) aponta que:

> A principal característica do processo de representação da informação é a substituição de uma entidade lingüística longa e complexa - o texto do documento - por sua descrição abreviada. O uso de tal sumarização não é apenas uma conseqüência de restrições práticas quanto ao volume de material a ser armazenado e recuperado. Essa sumarização é desejável pois sua função é demonstrar a essência do documento. Ela funciona então como um artifício para enfatizar o que é essencial no documento considerando sua recuperação, sendo a solução ideal para organização e uso da informação.

Na definição apresentada, verifica-se que o uso de instrumentos de representação da informação é essencial para que os documentos possam ser encontrados e representados para o processo de recuperação da informação. Como a própria autora destaca, a questão da representação está diretamente vinculada à recuperação, visto que esta última irá utilizar os elementos descritivos para que a busca possa ocorrer.

Outro autor que enfatizou a relação entre a representação e a recuperação da informação foi Souza (2006, p. 27), afirmando que:

> Tradicionalmente, as atividades de organização do conhecimento e representação da informação estiveram relacionadas a sistemas de recuperação de documentos. Os esquemas de classificação, gerais e especializados, os tesauros, entre outros tipos de instrumentos, foram criados para a organização física de acervos ou para a representação temática do conteúdo intelectual dos documentos visando acesso, disseminação e recuperação sistemática.

O trecho selecionado destaca que distintos instrumentos, como tesauros e esquemas, foram essenciais para que acontecesse a recuperação da informação dos acervos físicos, a princípio, e depois no formato digital. Todos esses instrumentos foram e continuam sendo essenciais dentro do contexto da recuperação da informação.

Vale apontar uma diferenciação existente entre os tipos de representação, sendo que ela pode ser temática ou descritiva. A representação temática preocupa-se em representar um determinado documento a partir dos assuntos que estão vinculados a ele, como, por exemplo, os conceitos que estão vinculados a uma determinada obra, enquanto a representação descritiva é focada em descrever elementos objetivos de uma obra, como título e autor.

Maiome et al. (2011, p. 27) destacam ainda que:

> A primeira [representação descritiva] representa as características específicas do documento, denominada descrição bibliográfica, que permite a individualização do documento. Ela também define e padroniza os pontos de acesso, responsáveis pela busca e recuperação da informação, assim como

> pela reunião de documentos semelhantes, por exemplo, todas as obras de um determinado autor ou de uma série específica. A segunda [representação temática] detém-se na representação dos assuntos dos documentos a fim de aproximá-los, tornando mais fácil a recuperação de materiais relevantes que dizem respeito a temas semelhantes. Neste contexto, são elaboradas as linguagens documentárias, instrumentos de controle vocabular a fim de tornar possível a "conversação" entre documentos e usuários.

As autoras destacam e apontam as diferenças entre os tipos de representação, além de demonstrar a importância de cada um deles. Vale apontar que, no âmbito da recuperação da informação, ambas as representações são fundamentais para que o processo de busca seja mais eficiente. A busca por características descritivas é essencial, tal qual a busca por assunto também é necessária.

Com a evolução das tecnologias e a possibilidade de análise de textos completos dos documentos foram criados mecanismos que, além de buscar em representações dos documentos, eram capazes de buscar nos textos, aprimorando assim os resultados encontrados. Em especial, os metadados são elementos de representação da informação que são amplamente utilizados no contexto bibliográfico e da Web, tornando-se essenciais para os atuais mecanismos de busca e recuperação da informação.

Metadados significa literalmente "dados sobre dados", ou seja, dados que descrevem outros dados. No entanto, essa definição é simples diante da evolução das tecnologias e dos estudos tratando da representação da informação e dos metadados.

Uma definição mais ampla foi dada por Alves (2010, p. 47):

> [...] atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação.

Vinculados ao conceito de metadados estão os padrões de metadados, que foram criados para formalizar e padronizar os metadados, de acordo com o cenário que está sendo descrito. Atualmente, existem diversos padrões, que vão desde o domínio bibliográfico, passando pelo arquivístico, e chegando até a descrição de elementos da Web.

No âmbito da Web, o padrão de metadados Dublin Core tem ganhado um destaque significativo, havendo uma ampla utilização desse padrão para descrição de objetos informacionais. O padrão Dublin Core é formado por quinze elementos, tendo como objetivo facilitar a identificação e a descoberta de recursos na Web, caracterizando-se como um padrão para propósitos gerais, que pode ser utilizado em diversos domínios.

Além de metadados, outros instrumentos como ontologias, tesauros e taxonomias podem ser utilizados para a realização do processo de recuperação da informação. Instrumentos que possuem um alto nível de semântica formal conseguem descrever com mais precisão, permitindo que a recuperação seja mais eficiente.

Nesse sentido, o presente trabalho parte dos pressupostos da representação da informação, em especial dos instrumentos oriundos das ferramentas da Web Semântica, que contribuem tanto para ter uma descrição mais detalhada e com mais significado, quanto para aprimorar a recuperação da informação.

Partindo desses pressupostos, apresenta-se, a seguir, a seção de Web Semântica, que traz os conceitos e as ferramentas desse campo de estudos.

# 3 WEB SEMÂNTICA

A Web Semântica vem se desenvolvendo nos últimos anos, tornando-se um elemento fundamental em diversos nichos de pesquisa, ao fornecer um arcabouço conceitual importante para o desenvolvimento da Web, um conjunto de ferramentas que contribuem para diversos contextos da Ciência da Informação e da Ciência da Computação e uma série de aplicações que estão aprimorando a forma como os usuários utilizam a Web e como os dados são nela publicados.

Dessa forma, a Web Semântica apresenta-se como o principal elemento do polo teórico deste trabalho, fornecendo subsídios teóricos e aplicados para que esta tese pudesse ser desenvolvida. Neste capítulo, serão abordadas as principais questões teóricas que foram a base para que o trabalho fosse desenvolvido. Para isso, esta seção está dividida em quatro subseções.

A primeira subseção apresenta a história da Web e da Web Semântica, demonstrando como a Web foi se desenvolvendo, fazendo com que a Web Semântica fosse algo inevitável. Além disso, apresentam-se os marcos para que a Web Semântica pudesse se tornar o que é hoje.

Na segunda subseção explicitam-se os principais conceitos que estão vinculados à Web Semântica, destacando-se, em especial, o conceito de disponibilização dos dados que a Web Semântica propõe e as ontologias, que se tornaram elementos essenciais para a construção da Web Semântica.

A terceira subseção destaca as ferramentas que foram desenvolvidas como uma consequência da Web Semântica.

Na última seção, apresenta-se o processo de materialização da Web Semântica e as aplicações que surgiram dentro desse contexto. Discute-se também o que hoje é considerado a principal aplicação para a materialização da Web Semântica, o *Linked Data*.

## 3.1 HISTÓRIA DA WEB E DA WEB SEMÂNTICA

A Web foi proposta em 1989 por Tim Berners-Lee, visando a ser um gerenciador de informações gerais sobre projetos desenvolvidos no *European Organization for Nuclear Research* (Organização Europeia para a Pesquisa Nuclear - CERN).

Originalmente, essa proposta foi construída por Berners-Lee com o objetivo de auxiliar no gerenciamento dos documentos do CERN, que apresentavam uma grande heterogeneidade de informações, e pessoas que trabalhavam em sedes espalhadas pela Europa necessitavam trocar dados com mais rapidez e com mais dinâmica. No texto que apresenta a proposta, Berners-Lee (1989) relata que o modo como as informações eram armazenadas e

compartilhadas no âmbito do CERN, em árvores, não era eficiente para os grandes projetos desta organização.

Diante desse cenário, o autor aponta que a solução seria utilizar o hipertexto como um meio para gerenciar as informações desses projetos do CERN. Conforme Berners-Lee (1989) aponta, o termo hipertexto foi cunhado por Ted Nelson na década de 1950. Destaca-se ainda que, nesse texto, Berners-Lee não chamava essa proposta de Web, mas sim de *Linked information systems* (sistemas de informações ligadas), justamente pela característica do hipertexto que permite a ligação dos recursos.

A partir dessa proposta, em 1990, Berners-Lee e Cailliau (1990), no texto denominado *WorldWideWeb: Proposal for a hypertext project*, explicam como esse projeto utilizaria o hipertexto como base para esse sistema de gerenciamento de informação. Nesse texto também o nome *WorldWideWeb* é cunhado, e a proposta começa a se tornar real e aplicável.

Os autores ainda apontam como o hipertexto está relacionado a essa proposta:

> O hipertexto é uma maneira de vincular e acessar informações de vários tipos como uma rede de nós nos quais o usuário pode navegar à vontade. Ele fornece uma única interface de usuário para grandes classes de informações (relatórios, notas, bases de dados, documentação do computador e ajuda on-line). Propomos um esquema simples incorporando servidores já disponíveis no CERN. (BERNERS-LEE; CAILLIAU, 1990, não paginado, tradução nossa)

A partir do que é relatado pelos autores, verifica-se que o hipertexto é central no contexto da Web, ao permitir a interligação dos dados e criação dos nós que caracteriza o hipertexto e a Web atual. Além disso, identifica-se que a proposta contemplou outras aplicações do CERN e começou a se materializar na forma que temos atualmente.

A partir desse texto do Berners-Lee e Cailliau (1990), o termo *World Wide Web*, que chamaremos somente de Web no âmbito deste trabalho, passou a ser uma iniciativa mundial, tornando-se a principal aplicação da internet. Em poucos anos, a Web se tornou muito utilizada, havendo diversas aplicações e páginas, chamadas de websites, fazendo parte desse projeto.

Uma característica da Web, essencial para compreender como essa iniciativa se desenvolveu, está vinculada à característica livre que a Web possui. No ano de 1993, a Web se tornou de domínio público, tendo como característica central ser descentralizada e livre de custos. Em outras palavras, a inserção de conteúdo na Web não seria controlada por nenhuma instituição ou organização, não necessitando haver o pagamento por isso.

Nesse sentido, Berners-Lee, em uma entrevista em 2003, relatou que:

> A decisão do CERN de tornar as bases e os protocolos da Web disponíveis sem royalties e sem impedimentos adicionais foi crucial para a existência da

> Web. Sem esse compromisso, o enorme investimento individual e corporativo em tecnologia da Web simplesmente nunca teria acontecido, e não teríamos a Web hoje. (EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH, 2003).

As palavras de Berners-Lee demonstram que a base da Web estava construída sob a liberdade, e a inserção dessa proposta em domínio público foi essencial para a Web se tornar o que é atualmente. Como relatado, essa característica tornou a Web descentralizada, permitindo que qualquer usuário pudesse inserir conteúdo nesse ambiente.

No entanto, essa descentralização da Web teve como consequência um aumento exponencial na quantidade de informações disponibilizadas. Nos anos iniciais da Web, informações dos mais diversos tipos foram sendo inseridas na Web, utilizando o hipertexto como base para possibilitar a interligação e a navegação pelos conteúdos. Porém, a grande maioria desses conteúdos foram estruturados e disponibilizados somente para a leitura humana. (SOUZA; ALVARENGA, 2004).

Quando se diz que uma informação está estruturada apenas para a leitura humana no contexto da Web, está vinculada, especialmente, a páginas Web que utilizam apenas elementos visuais das linguagens de hipertexto e linguagens de estilo, como o *HyperText Markup Language* (HTML) e *Cascading Style Sheets* (CSS). Uma consequência de apontar que uma página Web está preparada apenas para a leitura humana é que as informações estão estruturadas apenas visualmente, não fornecendo informações que expressem o significado das informações apresentadas, para eventuais mecanismos computacionais que busquem extrair conhecimento desses ambientes.

Nesse cenário, Souza e Alvarenga (2004, p. 133) apontam como a Web era vista em seus anos iniciais, em que não havia uma preocupação em expressar a semântica dos conteúdos para os mecanismos computacionais, pelas características que a Web possuía:

> Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a Web foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade.

Vale destacar que a visão dos autores expressa com clareza como a Web estava estruturada e organizada nos seus primeiros anos. Isso demonstra que o rápido crescimento da quantidade de conteúdos disponibilizados, consequência da revolução que esse ambiente provocou na sociedade, também trouxe dificuldades, que necessitavam, de alguma maneira, ser

tratadas, para não sentenciar a Web a ser finalizada por ser inviável navegar e encontrar quaisquer informações nesse ambiente.

Diante dessa dificuldade, Berners-Lee (1998) começa a pensar em uma forma de solucionar a problemática em que a Web estava envolta, em que havia muito conteúdo, mas sem significado algum para os mecanismos computacionais, inviabilizando a recuperação adequada e que permitisse a descoberta de novas informações. Nesse texto, o autor inicia as reflexões sobre uma Web Semântica, em que, a partir disso, esse termo passa a significar essa proposta onde a Web teria um nível semântico mais elevado.

Nessa seara, posteriormente, Berners-Lee, Hendler e Lassila (2001) começam a dar forma a Web Semântica como uma proposta para solucionar esse cenário, na busca de permitir que os conteúdos informacionais fossem providos com mais significados e mais informações descritivas. Assim, a proposta da Web Semântica traria: "[...] estrutura para o conteúdo significativo das páginas da Web, criando um ambiente no qual os agentes de software que transitam de uma página para outra possam realizar prontamente tarefas sofisticadas para os usuários." (BERNERS-LEE; HENDLER; LASSILA, 2001, não paginado).

Essa citação demonstra que o objetivo da Web Semântica sempre foi facilitar as tarefas realizadas pelos usuários, em especial pela dificuldade existente naquele momento para o usuário realizar até mesmo tarefas simples. Além disso, destaca-se que a estrutura em que as informações estavam disponibilizadas, não permitiria aos mecanismos computacionais expandirem o modo como os usuários navegam e utilizam a Web.

Um outro ponto central na compreensão do papel da Web Semântica está na sua relação com a Web. Tal relação é explicitada já no texto inicial da Web Semântica de 2001, em que Berners-Lee, Hendler e Lassila (2001, não paginado, tradução nossa) afirmam que:

> A Web Semântica não é uma Web separada, mas uma extensão da atual, na qual a informação possui um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação. Os primeiros passos para tecer a Web Semântica na estrutura da Web existente já estão em andamento. Em um futuro próximo, esses desenvolvimentos trarão novas funcionalidades significativas, à medida que as máquinas se tornarem muito mais capazes de processar e "entender" os dados que elas simplesmente exibem no momento.

Duas questões podem ser destacadas do trecho citado. A primeira está no apontamento da Web Semântica ser uma extensão da Web, em que os dados dessa última teriam um nível maior de formalização semântica. Tal questão demonstra que a Web Semântica foi proposta com o intuito de aprimorar a Web, permitindo que esse ambiente tivesse mais possibilidades para aprimorar a forma como os usuários interagem e acessam as informações.

A segunda questão está no processo de desenvolvimento da Web Semântica, em que os autores demonstram como ocorreu esse processo, em que as ferramentas foram construídas a partir dessas bases teóricas apresentadas por eles e pelos demais pesquisadores que buscavam aprimorar a Web.

Nesse âmbito, Ramalho, Vidotti e Fujita (2008, não paginado) apontam que:

> [...] observa-se que comparando com as abordagens tradicionalmente desenvolvidas, o projeto Web Semântica constitui-se como uma tentativa inversa de solução que tem como objetivo desenvolver meios para que as máquinas possam servir aos humanos de maneira mais eficiente, mas para isso torna-se necessário construir instrumentos que forneçam sentido lógico e semântico aos computadores

A questão abordada pelos autores expressa com clareza os objetivos da Web Semântica, bem como qual caminho essa proposta seguiu para solucionar a questão expressa da problemática da falta de organização existente na Web, que conduzia a uma ineficiência em alguns processos. A inversão que os autores indicaram demonstra que a Web Semântica buscou expandir a compreensão dos conteúdos pela máquina, para que a máquina servisse melhor aos indivíduos. Nesse sentido, como relatado anteriormente, a proposta esteve sempre focada em construir um ambiente melhor para as pessoas.

A partir dessa concepção inicial apresentada por Berners-Lee, Hendler e Lassila (2001), a Web Semântica começou a se desenvolver de modo a permitir que essa proposta se tornasse de fato real e implementável. Assim, nos anos seguintes a 2001, uma série de novas tecnologias e ferramentas começaram a ser desenvolvidas, tornando a Web Semântica mais próxima da realidade, influenciando o modo como as novas aplicações passaram a ser construídas e permitindo uma comunicação com novas tecnologias de outros campos de estudos, como a Inteligência Artificial.

Nesse contexto, cinco anos após a proposta inicial da Web Semântica, Shadbolt, Hall e Berners-Lee (2006) escreveram o texto *The Semantic Web Revisited*, que evidencia como a Web Semântica amadureceu nesses primeiros anos, demonstrando como as ferramentas da Web Semântica se tornaram realidade, ao mesmo tempo que indicava a necessidade de haver uma maior integração entre tais tecnologias, a Web e a Inteligência Artificial. Nesse sentido, os autores apontam que:

> Esperamos que os desenvolvimentos, metodologias, desafios e técnicas que discutimos aqui não apenas originem uma Web Semântica, mas também contribuam para uma nova Ciência da Web - uma ciência que busca desenvolver, implantar e compreender sistemas e sistemas de informação distribuídos, de seres humanos e máquinas, operando em escala global. IA [Inteligência Artificial] será uma das disciplinas contribuintes. IA já nos

> forneceu métodos de programação funcional e lógica, formas de compreender sistemas distribuídos, ferramentas de detecção de padrões e de mineração de dados, abordagens de inferência, engenharia ontológica e representação de conhecimento. Tudo isso é fundamental para uma agenda da Web Science e para a realização da Web Semântica. (SHADBOLT; HALL; BERNERS-LEE, 2006, p. 101, tradução nossa)

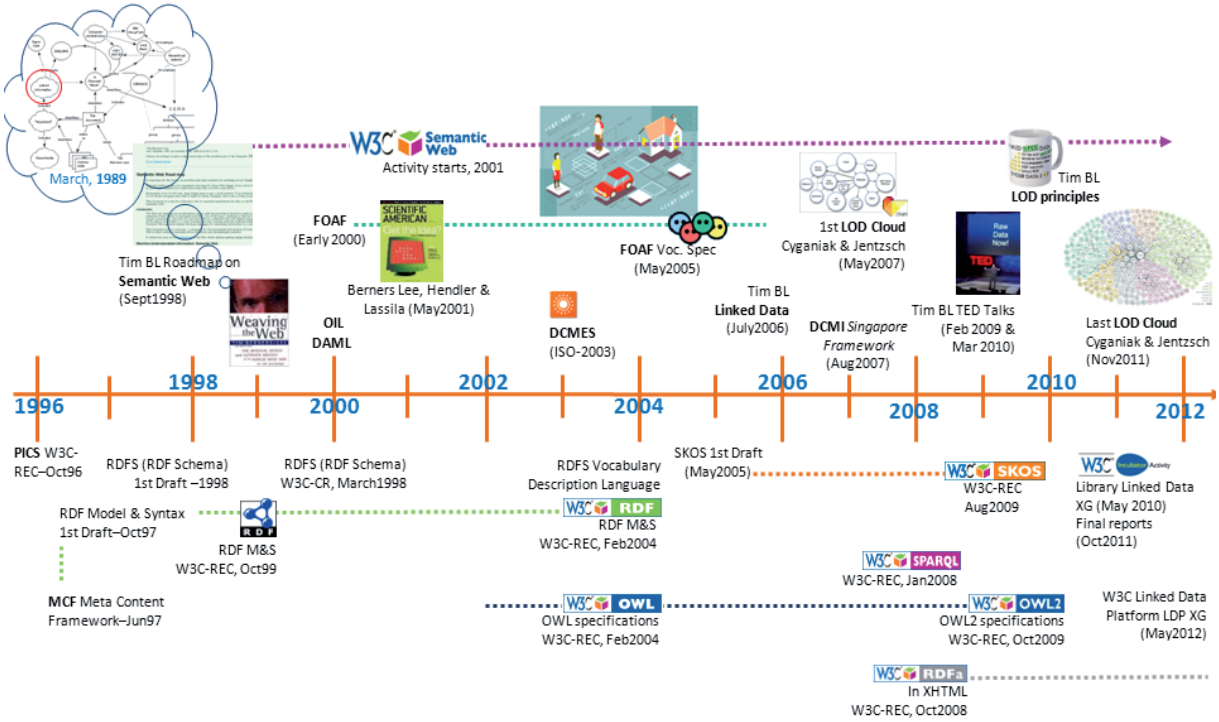
O trecho apresentado relata com clareza como a Web Semântica está vinculada e pode contribuir com outros campos de estudos, como a própria Inteligência Artificial. Destaca-se que a materialização da Web Semântica é essencial para fornecer subsídios para essas tecnologias distintas, pois é a primeira que deve ser responsável no fornecimento de informações, com um nível de semântica formal adequado para a realização de inferências e a descoberta de novas informações, apoiadas nas aplicações de Inteligência Artificial, entre outros campos de estudos da Ciência da Computação.

Nesse momento em que o texto supracitado foi escrito, a proposta do *Linked Data* iniciava, de forma mais contundente, o processo chamado de materialização da Web Semântica. O termo materialização da Web Semântica foi discutido por diversos autores, entre eles: Arantes (2010), Santarem Segundo (2015), Santarem Segundo e Coneglian (2016) e Coneglian (2017).

Vale destacar que os três últimos textos citados apontam ainda que o *Linked Data* se mostra como a principal iniciativa de materialização da Web Semântica, como relatado por Santarem Segundo (2015, p. 225) ao apontar o *Linked Data*: “[...] como a melhor forma de materialização dos conceitos e tecnologias da Web Semântica [...]”.

A proposta do *Linked Data* utiliza as tecnologias e as ferramentas da Web Semântica para interligar os dados na Web, permitindo que os princípios destacados no início da Web Semântica fossem implementados. Essa proposta será discutida com mais detalhes na subseção 3.4, porém a figura 3, a seguir, já permite visualizar como a Web Semântica foi evoluindo de modo a tornar possível a existência do *Linked Data*.

**Figura 3** - A evolução da Web Semântica até o *Linked Data*




Fonte: Méndez e Greenberg (2012, p. 238).

O levantamento realizado por Méndez e Greenberg (2012), apresentado na figura 3, aponta, na parte superior, alguns dos principais textos que marcaram a história da Web Semântica até o ano de 2012, enquanto na parte inferior se observam as principais ferramentas da Web Semântica, que conduzem até o momento em que o *Linked Data* se encontrava no ano do estudo.

Destaca-se que há uma íntima relação entre o modo como os textos foram construídos, as tecnologias se desenvolveram e a forma como o *Linked Data* se tornou viável. Atualmente, o *Linked Data* é um projeto que tem se expandido significativamente, tornando-se um importante meio para a publicação de dados de diversas áreas de estudos, como Ciências da Saúde, Ciências naturais e dados governamentais.

No momento atual, a Web Semântica encontra-se em um momento de expansão para campos diversos, com diversas aplicações e com os mais distintos objetivos. As ferramentas e os conceitos da Web Semântica estão presentes em projetos utilizados por milhões de pessoas por dia, além de haver uma manutenção e fortalecimento do *Linked Data*.

Dentre os projetos que estão baseados nos princípios da Web Semântica, destaca-se uma patente da empresa de mídias sociais Facebook, em que a forma como se dá a busca para a

localização das entidades e dos recursos na rede social, utilizando-se dos conceitos de grafos e informações estruturadas, tem uma relação direta com a proposta da Web Semântica (FACEBOOK, 2016). Apesar de na patente não haver nenhuma indicação da Web Semântica, os princípios dessa proposta estão claramente apresentados na forma como a patente se coloca.

Outro projeto de grande alcance é o Google *Knowledge Graph.* Singhal (2012), que, ao apresentar essa proposta, explicita que a ideia do projeto é transformar o modo como a busca é realizada, o que se evidenciaria ao tratar as informações como objetos e não somente cadeias de caracteres a ser buscadas. A relação dessa proposta com a Web Semântica foi tratada por diversos autores como Monteiro (2015), Santarem Segundo, Souza e Coneglian (2015) e Coneglian et al. (2017).

A forma como o *Knowledge Graph* utiliza os princípios da Web Semântica foi verificada por Coneglian et al. (2017, p. 46), ao afirmar que é possível: "[...] identificar que a estrutura semântica de todo o *Knowledge Graph* ocorre inteiramente baseada em RDF, em que as relações são sempre, sujeito, predicado e objeto [...]" (CONEGLIAN et al., 2017).

Um outro projeto de destaque internacional baseado nos princípios da Web Semântica e do *Linked Data* é a Europeana. A Europeana reúne documentos de bibliotecas, arquivos e museus dos diversos países da Europa, sendo um importante sistema para a preservação da memória do continente. Nesse sentido, as informações contidas nesse ambiente estão estruturadas seguindo o *Europeana Data Model* (EDM), que foi estruturado utilizando as ferramentas da Web Semântica e é uma importante base de dados do *Linked Data*. (EUROPEANA, 2014)

Há outros projetos que estão sendo desenvolvidos ou que já foram finalizados, demonstrando o quanto a Web Semântica se expandiu e hoje é um importante elemento a ser considerado, para se refletir como as próximas ferramentas irão se desenvolver e influenciarão a vida dos indivíduos.

Todo o histórico e os conceitos apresentados possibilitam uma compreensão de como a Web Semântica evoluiu e se materializou, permitindo atualmente aos pesquisadores refletirem e desenvolverem elementos que extrapolam a própria Web, e que permitem utilizar os conceitos e as ferramentas dessa proposta para aprimorarem outros campos de estudos.

O exemplo mais evidente está na Inteligência Artificial, em que há uma interseção em questões sobre inferências, permitindo aprimorar a compreensão dos sistemas informacionais sobre as informações produzidas pelos usuários, retornando melhores resultados para esses mesmos usuários. Além disso, há ferramentas como motores de inferência, que estão cada vez

mais avançando na busca de melhorar significativamente as consultas e a realização de relacionamentos entre os dados que estão disponibilizados nas ferramentas da Web Semântica.

No entanto, essa relação com outros campos de estudos não fica restrito à Inteligência Artificial, podendo ser expandido para disciplinas que podem aprimorar ainda mais essa relação entre os mecanismos computacionais e os indivíduos. Em especial, o processamento de linguagem natural, que há décadas vem sendo estudado e trabalhado nas áreas da Ciência da Computação, pode fornecer subsídios importantes para a Web Semântica, ao mesmo tempo que pode receber contribuições essenciais, para tornar mais clara a compreensão da forma como os usuários se comunicam.

No contexto deste trabalho, a aproximação desses dois campos de estudos, Web Semântica e processamento de linguagem natural, visa a avançar a forma como a comunicação entre humanos e máquinas se comunicam, inserindo a compreensão do significado e do contexto aos ambientes informacionais digitais. Nesse sentido, a Ciência da Informação é a base epistemológica e teórica para permitir que essas relações entres os campos ocorra, tendo como ponto central o indivíduo, elemento primordial para permitir a compreensão do contexto em que os usuários se encontram.

Essa abordagem final demonstra o que esse trabalho visa a trazer para continuar essa linha temporal da Web Semântica, apresentando a coerência ao modo como esse campo de estudos foi evoluindo tanto na Ciência da Computação, quanto na Ciência da Informação.

A forma como a Web Semântica evoluiu está vinculada necessariamente a alguns conceitos que nasceram a partir da Web Semântica, ou que já existiam anteriormente, mas de algum modo foram vinculados fortemente aos estudos desse campo. Os conceitos da Web Semântica são um ponto-chave para compreender como esse trabalho será desenvolvido e os elementos que compõem o modelo proposto. Assim, na sequência, apresentam-se os principais conceitos vinculados aos estudos de Web Semântica.

## 3.2 CONCEITOS DA WEB SEMÂNTICA

A divisão da Web Semântica em conceitos e ferramentas possibilita uma compreensão mais clara do impacto que esse campo de estudo teve em diversas áreas do conhecimento. Isso ocorre, pois os conceitos da Web Semântica estão sendo largamente utilizados em diversas áreas do conhecimento. Esses conceitos permitem a construção de modelos teóricos e embasam o desenvolvimento de ferramentas e aplicações.

Nesse sentido, nesta subseção, serão discutidos os principais conceitos da Web Semântica que embasam os estudos aqui realizados: a representação de recursos, tratado de modelo RDF, e as ontologias.

### 3.2.1 Representação dos recursos (RDF)

O primeiro conceito central da Web Semântica e que tem impactado o desenvolvimento de diversos outros ferramentais e tecnologias é o modo como a representação dos recursos é feita por meio do RDF.

Primeiramente, vale ressaltar que o RDF é tratado como um modelo de representação de recursos e, por vezes, é considerado uma tecnologia da Web Semântica, pelo seu uso nas mais diversas ferramentas que estão dentro do escopo desse campo de estudo. No entanto, o conceito que embasa a representação de recursos utilizando o modelo RDF tem extrapolado o próprio RDF, uma vez que há uma série de aplicações, que tem em sua base os conceitos da Web Semântica, que utilizam o princípio do RDF, sem utilizar o modelo em si.

Destacam-se dois projetos principais: o *Knowledge Graph*, do Google, e patentes do Facebook que usam o conceito do RDF. Ambos os projetos não citam diretamente a Web Semântica e, mais especificamente, não citam o RDF como tecnologia base para o seu desenvolvimento.

Como citado anteriormente, o primeiro projeto, o *Knowledge Graph* é uma expansão do modo como o Google realiza as suas buscas, por meio de associações realizadas entre os recursos. O segundo, referente a uma patente do Facebook, trata de como as buscas, dentro dessa rede social, ocorrem por meio de associações e ligações existentes entre as diversas informações que essa ferramenta possui. (FACEBOOK, 2016; SINGHAL, 2012).

Esses dois projetos utilizam em sua essência o princípio do RDF, que está na associação dos recursos. Esse modelo busca tornar o relacionamento e o armazenamento mais associativo, onde a ligação entre os recursos ocorre sempre por meio de algum atributo.

> RDF é um modelo padrão para intercâmbio de dados na Web. O RDF possui recursos que facilitam a fusão de dados, mesmo se os esquemas relacionados forem diferentes, e suporta especificamente a evolução dos esquemas ao longo do tempo, sem exigir que todos os consumidores de dados sejam alterados. O RDF estende a estrutura de vinculação da Web para usar URIs para nomear a relação entre as coisas, bem como as duas extremidades do link (isso geralmente é chamado de "tripla"). Usando esse modelo simples, ele permite que dados estruturados e semiestruturados sejam misturados, expostos e compartilhados em diferentes aplicativos. Essa estrutura de vinculação forma um gráfico rotulado e direcionado, no qual as arestas representam o link nomeado entre dois recursos, representados pelos nós do gráfico. Essa visão

> gráfica é o modelo mental mais fácil possível para o RDF e é frequentemente usada em explicações visuais de fácil compreensão. (WORLD WIDE CONSORTIUM, 2014, tradução nossa)

A análise desse trecho pode ser realizada sob uma perspectiva conceitual, em que o RDF é visto como um conceito da Web Semântica, e não como uma tecnologia. Tendo esse prisma, verifica-se que o RDF aprimora o modo como as informações podem ser recuperadas, armazenadas e associadas. Destaca-se que o relacionamento dos dados por meio de grafos permite que, tanto dados estruturados, quanto dados não estruturados, estejam interligados.

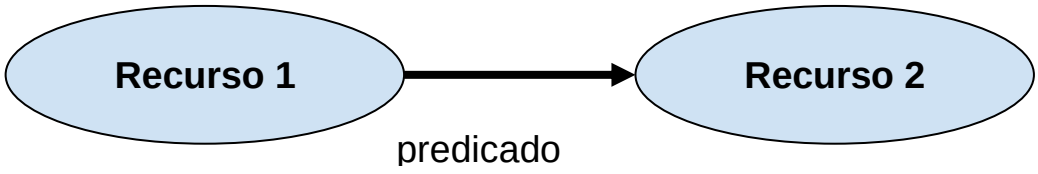
Nesse contexto, o uso do princípio do RDF permite que vídeos estejam associados a textos e a outras mídias, de modo que a recuperação da informação seja muito mais simples e natural, tanto para os mecanismos computacionais, quanto para os indivíduos.

Esse modelo se contrapõe a outros modos em que as informações são relacionadas e armazenadas, no qual se destaca o modelo relacional, utilizado pelos principais sistemas gerenciadores de banco de dados. O modelo relacional organiza as informações por meio de tabelas, de forma que, em uma tabela, as linhas são registros e as colunas contêm as informações de cada registro. Outro importante aspecto do modelo relacional, é que é possível realizar relacionamento entre os registros, utilizando identificadores de uma tabela, fazendo referência a outras tabelas.

Em contraposição ao modelo relacional, considera-se que o uso do conceito do RDF para o relacionamento e o armazenamento de recursos se embasa em outros princípios, em que todos os dados estão relacionados e são armazenados por meio de ligações.

Um conceito base para entender o funcionamento do RDF está no conceito de modelo de dados em grafos, que pode ser observado na figura 4.

**Figura 4** - Princípio do relacionamento RDF




Fonte: elaborado pelo autor.

O princípio apresentado na figura 4 está no relacionamento entre dois recursos por meio de algum predicado, que vincula duas coisas. Para compreender melhor o funcionamento desse princípio de relacionamento do RDF, pode-se entender que o RDF é composto por uma tripla

RDF, em que o sujeito (recurso 1, na figura 4) está ligado por meio de uma propriedade (predicado, na figura 4) a um objeto (recurso 2, na figura 4). Isso é validado por Klyne e Carroll (2004, tradução nossa) que, no documento que apresenta o RDF oficialmente, apontam: "A afirmação de um triplo RDF diz que alguma relação, indicada pelo predicado, se mantém entre as coisas denotadas por sujeito e objeto do triplo."

Em suma, o sujeito indica uma informação, que pode ser desde um artigo científico, a um autor, ou mesmo uma música. Já o predicado é "[...] um recurso que possui um nome e pode ser utilizado para caracterizar um outro recurso, como, por exemplo, criador e título" (BREITMAN, 2005, p. 22). Santarem Segundo (2014, p. 3866) complementa relatando que "[...] são os atributos que permitem distinguir um recurso de outro ou que descrevem o relacionamento entre recursos". E o objeto são os dados que representam o conteúdo que está sendo descrito, que podem ser informações literais ou outros recursos, como relatado por Breitman (2005).

O próximo princípio que apresenta e valida o RDF está vinculado à necessidade de os nós terem identificação e serem parte de vocabulários baseados em URI. Basicamente isso vem pela necessidade dos recursos e objetos serem identificáveis unicamente, além de possibilitar que estes possam ser encontrados pela Web.

Destaca-se que "Um *Uniform Resource Identifier* (URI) fornece um meio simples e extensível para a identificação de um recurso." (BERNERS-LEE, et al., 2005, tradução nossa). Coneglian e Santarem Segundo (2017, p. 89-90) complementam essa afirmação, ao considerar que "Um URI apresenta importância devido à necessidade da identificação fácil dos recursos da Web. Outra característica relevante é a uniformidade dos recursos, que promove a diferenciação dos mesmos, facilitando a compreensão do contexto que um recurso possui."

Adicionalmente, Ferreira e Santos (2013, p. 18) destacam que "O importante, no entanto, não é a recuperação ou não de algo por um navegador a partir desse URI, nem mesmo se o que é recuperado tem ou não alguma relação com o livro em questão, mas sim a própria identificação do recurso."

Assim, verifica-se que o RDF passa a ter um papel central na interoperabilidade e na identificação das informações, o que pode ser verificado pela exigência do uso de URIs. Ferreira e Santos (2013, p. 21) continuam ainda: "O modelo RDF oferece a possibilidade para as comunidades de descrição de recursos definirem a semântica de seus metadados de maneira formal, isto é, definindo o significado dos elementos de metadados, conforme as suas necessidades específicas de descrição, em um modelo processável por máquinas."

O terceiro conceito que pauta o RDF está nos tipos de dados que podem ser utilizados nos dados modelados com RDF. Em suma, um tipo de dado se trata de um "[...] espaço léxico, um espaço com valores e um mapeamento léxico para um valor." (KLYNE; CARROLL, 2004, tradução nossa). Esses tipos de dados são utilizados para representar valores inteiros, *strings*, números reais, booleanos, entre outros. Além disso, o próprio RDF permite a criação de novos tipos de dados, dando assim flexibilidade para o processo de representação.

Vinculado ao conceito dos tipos de dados, tem-se o quarto princípio do RDF que são os valores literais e não-literais. Os valores literais são valores absolutos, como *strings* ou números, enquanto, os valores não-literais são os recursos.

Ferreira e Santos (2013, p. 19) aprofundam essa questão, relatando que:

> Literais são valores de dados de um certo tipo de dados (*datatype*). O valor de cada literal é geralmente descrito como uma sequência de caracteres, tais como a cadeia de caracteres composta pelos símbolos 3 e 6 do exemplo anterior. A interpretação de tais sequências realizada pela máquina é, então, baseada em um tipo de dados específico.

Dessa forma, todos os recursos e valores que são constituídos por recursos que possuem URI são chamados de valores não-literais, isso porque fazem referências a conceitos e objetos, que não são apenas constituídos por uma única informação. Isso é importante, pois é uma das principais características do RDF, permitindo que um conjunto de dados em uma mesma estrutura possua dados literais e não-literais.

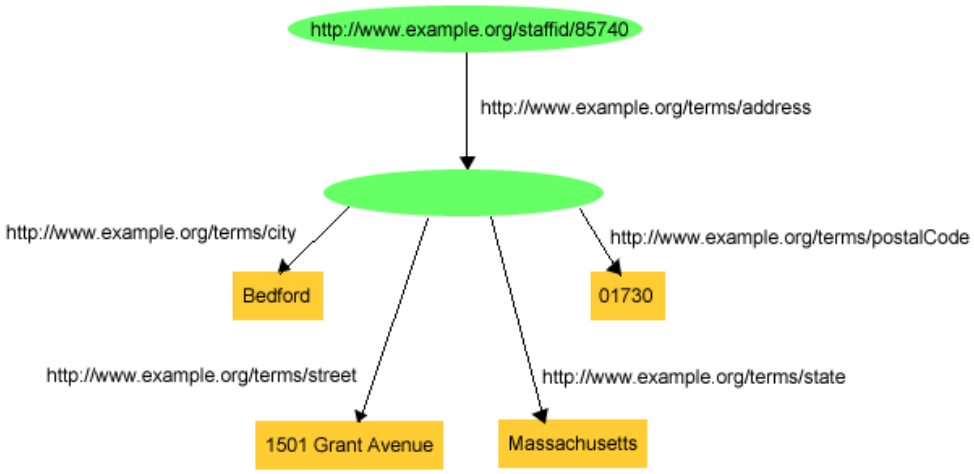
Por fim, o último conceito importante do RDF está no modo como um simples fato pode ser representado. Em suma, quando se representam informações em que um recurso precisa estar conectado a diversos outros recursos, o RDF utiliza de um conceito de nó branco para possibilitar que esse relacionamento ocorra. Klyne e Carroll (2004, tradução nossa) apontam que: "Uma forma simples de decomposição introduz um novo nó em branco, correspondente à linha, e uma novo tripla é introduzida para cada célula na linha. O sujeito de cada tripla é o novo nó em branco, o predicado corresponde ao nome da coluna e o objeto corresponde ao valor na célula."

Coneglian (2017, p. 79) exemplifica essa situação apontando que:

> Um exemplo possível de existência de nós em branco, é a relação de pessoa com endereço. A tripla RDF desse exemplo seria composto por: a pessoa como sujeito, o relacionamento como endereço, e o objeto seria o valor que informa o endereço; entretanto, como o endereço possui diversos dados, como rua, número, bairro, cidade, estado, o RDF permite que o objeto seja um nó em branco, que se relaciona com diversos outros nós; nesse exemplo, em particular, o nó em branco se relaciona com rua, número, bairro, cidade e estado

A figura 5 demonstra o que foi relatado pelo autor e o mecanismo chamado de nó branco.

**Figura 5** - Exemplo de grafo com nó em branco




Fonte: Klyne e Carroll (2004)

Os conceitos apresentados e a explanação realizada estão focados no RDF enquanto conceito, e não em seu uso como tecnologia. Posteriormente, na seção 3.3.1 será apresentado como o RDF pode ser utilizado como ferramentas, com XML e outras tecnologias.

A seguir apresenta-se o conceito de ontologias, elemento central para este trabalho.

### 3.2.2 Ontologias

As ontologias estão se tornando importantes elementos da Web Semântica, para possibilitar informações capazes de contextualizar um cenário, bem como para fornecer informações que possam ser compreendidas, computacionalmente, no que diz respeito ao sentido que determinados termos podem ter. Nesse contexto, desde o início da Web Semântica, as ontologias estão se tornando cada vez mais populares e utilizadas em diversos cenários, havendo diversas iniciativas que buscam construir ontologias para fornecer uma descrição formalizada de um domínio.

Destaca-se que o termo ontologia é utilizado para fazer referência a um ramo da filosofia desde os séculos XVII e XVIII, como apontado por Ramalho (2006). No entanto, Coneglian (2017) destaca que, mais recentemente, a partir da década de 1990, o termo ontologia passou a ser utilizado em pesquisas das áreas de Ciência da Informação e Ciência da Computação, tendo um foco diferente daquele da filosofia, fazendo referência a instrumentos que buscam demonstrar o contexto de um cenário existente no mundo real para o computador.

A primeira definição dada neste trabalho é uma das mais populares encontradas na literatura, e é feita por Gruber (1993, p.1, tradução nossa), apontando que: “[...] uma ontologia é uma especificação explícita de uma conceitualização [...]”. O autor, ao apontar especificação explícita indica a existência de conceitos e relacionamentos entre classes, que é apresentada por um conceitualização, que indica a ocorrência desse fato por meio de um modelo de algum domínio específico.

Essa definição é complementada por outro conceito bastante utilizado, que foi criado por Borst (1997), destacando que: “[...] uma ontologia é uma especificação formal de uma conceitualização compartilhada.” Destacam-se dois novos elementos nessa definição, o primeiro, a conceitualização, é vinculado à questão de ser uma especificação formal, o que indica a necessidade da ontologia ser capaz de ser processável e compreendida pelo computador, enquanto o segundo, o ser compartilhada, demonstra que ontologias devem ser criadas em comunidades, devendo refletir um conhecimento que é consensual.

Essas duas definições são as mais utilizadas na literatura, sendo fundamentais para compreender como as ontologias foram concebidas em seu início, e como esses instrumentos foram essenciais e influenciaram a proposta da Web Semântica.

Posteriormente, Guarino (1998, p. 5, tradução nossa) trouxe uma nova definição, relacionada às anteriores, mas trazendo novos elementos: “Uma ontologia é uma teoria lógica que representa o significado pretendido de um vocabulário formal.” Tal definição destaca que as ontologias representam vocabulários, trazendo a questão de ser uma forma de representar um determinado cenário.

Ramalho (2006) complementa tal questão apontando que:

> De acordo com tais considerações, uma ontologia é uma teoria lógica cujo modelo restringe uma conceitualização particular, sem especificar exatamente qual, ou, em outras palavras, pode-se definir como uma caracterização axiomática do significado de um vocabulário lógico, a qual tem o compromisso apenas com a consistência em um determinado domínio, e não com a completude.

A visão do autor relaciona as diversas definições clássicas, trazendo detalhes sobre o modo como as ontologias são capazes de possibilitar os elementos destacados nas definições. Em especial, pontos como os axiomas e a consistência são levantados como importantes para permitir a conceitualização de um domínio, permitindo criar vocabulários que trazem características lógicas.

Trazendo uma visão mais clara de como as ontologias estão relacionadas às áreas da Ciência da Informação e da Ciência da Computação, Campos e Campos (2014, p. 3827-3828) apontam que as ontologias fornecem: “[...] um modelo para representar os pressupostos

epistemológicos e ontológicos, relevantes para o entendimento de pesquisas e seu tratamento computacional através das iniciativas de dados interligados abertos, mas sua elaboração é um processo custoso."

As autoras destacam que as ontologias possuem uma alta expressividade e capacidade de representação, ao mesmo tempo que apresentam que as ontologias podem ser utilizadas nas iniciativas de dados abertos interligados. Essa característica será melhor explorada em uma subseção posterior sobre *Linked Data*.

Em uma visão mais aplicada à Web Semântica, Santarem Segundo e Coneglian (2015, p. 227) complementam o conceito de ontologias, do ponto de vista de aplicação tecnológica, dizendo que "[...] entendem-se as ontologias como artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de classes, propriedades, relações, restrições, axiomas e instâncias."

A visão dos autores aponta as ontologias de forma mais aplicada, mas trazendo o conceito que embasa essa visão, de elementos que descrevem um determinado domínio. Adicionalmente, os autores trazem elementos que fazem parte de uma ontologia, apontando como elas podem efetivamente cumprir sua função de especificar um contexto.

Nessas visões, verifica-se como o conceito de ontologia foi evoluindo com o passar dos anos e como a Web Semântica contribuiu para transformar e aprimorar a sua concepção. Nesse sentido, nas primeiras visões, de Gruber (1993), Borst (1997) e Guarino (1998), as ontologias eram vistas como instrumentos focados na contextualização de um domínio, tendo uma perspectiva mais teórica, sem trazer com tanta ênfase a questão computacional, ao passo que Ramalho (2006), Campos e Campos (2014) e Santarem Segundo e Coneglian (2015) começaram a mostrar ontologias com características mais aplicadas e computacionais. Essa reflexão aponta como as ontologias passaram a ser efetivamente instrumentos de fundamental importância para a Web Semântica e outras áreas.

Outra importante classificação acerca das ontologias está na forma de uso que elas podem ter. Essa classificação foi realizada por Guarino (1998) e divide as ontologias em quatro classes:

- Ontologias de topo (*top-level ontologies*): descrevem conceitos gerais e amplos, como espaço, tempo, matéria, objeto, evento, ação, entre outros. Destaca-se que tais ontologias não devem estar ligadas a um domínio ou problema específico, pois são generalistas.
- Ontologias de domínio (*domain ontologies*): ontologias mais específicas que as de topo, e buscam descrever um determinado domínio particular, podendo, inclusive,

especializar conceitos oriundos de ontologias de topo. Como exemplo delas podem-se considerar as ontologias da área da medicina.

- Ontologias de tarefa (*task ontologies*): ontologias que tratam de uma tarefa dentro de um domínio, sendo, portanto, especificações de ontologias de domínio. Exemplos seriam ontologias tratando de prontuários médicos dentro de um domínio de medicina.
- Ontologias de aplicação (*application ontologies*): ontologias que descrevem uma aplicação, dependendo tanto de um domínio quanto de uma tarefa, sendo por vezes uma especificação de ontologias de ambos os tipos (ontologias de domínio e de tarefa). Algumas ontologias desse tipo possuem conceitos que correspondem ao papel desempenhado por algum ator dentro de um domínio, durante a execução de uma determinada tarefa.

A classificação apresentada demonstra que as ontologias podem estar inseridas em diversos contextos, e com diferentes abrangências. No entanto, há uma complementaridade entre elas, de modo que, por vezes, uma ontologia é construída sendo uma especificação ou uma generalização de uma outra ontologia existente.

No âmbito da Web Semântica, a ontologia sempre foi considerada um dos principais elementos para a efetivação dessa proposta. No texto inicial da Web Semântica, de 2001, Berners-Lee, Hendler e Lassila (2001, tradução nossa), ainda sob uma perspectiva da Web, relatam que:

> As ontologias podem melhorar o funcionamento da Web de várias maneiras. Elas podem ser usadas de maneira simples para melhorar a precisão das pesquisas na Web - o programa de pesquisa pode procurar apenas as páginas que se referem a um conceito preciso, em vez de todas as que usam palavras-chave ambíguas. Aplicativos mais avançados usarão ontologias para relacionar as informações em uma página com as estruturas de conhecimento e regras de inferência associadas.

A explanação feita pelos criadores da Web Semântica revela o poder e a importância que as ontologias teriam para a proposta, mostrando que as ontologias se tornaram cada vez mais necessárias para a realização de processos complexos e fundamentais, como a realização de inferências.

Nesse contexto, Santarem Segundo e Coneglian (2016, p. 240) apontam as necessidades para a realização de axiomas para as inferências:

> Para que as inferências sejam possíveis do ponto de vista computacional, é necessário que sejam construídos regras e axiomas que conduzam os agentes computacionais a tomarem decisões acerca de informações que não estão

> explicitadas em conjuntos de informações, mas que certamente são passíveis de dedução.

As ontologias são os únicos instrumentos de representação que possuem esse nível de expressividade, a ponto de possibilitar a realização de inferências, como destacam os autores no texto.

Nesse sentido, aponta-se que há diversos tipos de vocabulários, como relatado por Lassila e McGuiness (2001), com as seguintes categorias: 1) Vocabulários controlados/catálogos; 2) Termos/glossário; 3) Tesauros; 4) Hierarquias tipo-de informais; 5) Hierarquias tipo-de formais; 6) Frames (propriedades); 7) Restrições de valores; 8) Restrições lógicas (disjunção, inverso, parte de), em que o primeiro apresenta a menor representatividade semântica e o último, a maior.

Essa classificação é importante para demonstrar que as ontologias são os únicos instrumentos que apresentam todos esses elementos, sendo, assim, os mais completos e apropriados para a realização de inferências. Coneglian (2017, p. 50) aponta ainda que: "As ontologias, além de apresentar relações hierárquicas e os relacionamentos semânticos que um tesauro possui, dispõem de restrições de valores e permitem a inserção de lógica na definição da semântica [...]".

No contexto da Web Semântica, a linguagem OWL é capaz de aplicar e de assumir a expressividade semântica relatada pelos autores. Isso porque, como afirma Santarem Segundo e Coneglian (2016, p. 221), "Para que se tornem efetivamente computacionais, as ontologias precisam passar de uma estrutura conceitual para implementação através de uma linguagem. Há um conjunto de linguagens que foram desenvolvidas ao longo dos anos para representação."

A linguagem OWL e outras ferramentas da Web Semântica que foram concebidas a partir de alguns dos conceitos relatados na presente subseção serão apresentadas a seguir.

## 3.3 FERRAMENTAS DA WEB SEMÂNTICA

As ferramentas da Web Semântica estão sendo desenvolvidas e aprimoradas significativamente desde a sua proposta inicial em 2001. Atualmente, as ferramentas da Web Semântica estão sendo utilizadas em diversos âmbitos, não estando restritas ao domínio da Web.

As áreas de recuperação da informação, organização da informação e Inteligência Artificial estão utilizando e contribuindo no aprimoramento das principais ferramentas da Web

Semântica. Nesta subseção serão apresentadas e discutidas as principais ferramentas da Web Semântica que estão interligados à pesquisa.

### 3.3.1 RDF/XML, N-Triple, Turtle e JSON LD

A tecnologia do RDF é uma das mais utilizadas na Web Semântica, pois os dados publicados seguindo os princípios dessa proposta são sempre disponibilizados nesse modelo. No entanto, o conceito do RDF pode ser efetivado por meio de diversos tipos de ferramentas, como XML, JSON, Turtle, entre outros. Destaca-se que isso ocorre devido à evolução tecnológica e a necessidades diversas que existem nas aplicações, que exigem formatos distintos.

A forma mais tradicional de utilizar e disponibilizar os dados de RDF é por meio do XML. Bray et al. (2006) apontam que o padrão XML foi criado em 1996, patrocinado pelo W3C, por meio do grupo de trabalho *XML Working Group*, liderado por Jon Bosak, da empresa Sun Microsystems (BRAY et al., 2006).

A *World Wide Consortium* (2011) descreve o XML como:

> [...] formato de texto derivado de SGML (ISO 8879) simples e muito flexível. Originalmente concebido para enfrentar os desafios da publicação eletrônica em grande escala, XML também está desempenhando um papel cada vez mais importante na troca de uma ampla variedade de dados na Web e em outros lugares.

O XML foi apontado por Berners-Lee, Hendler e Lassila (2001) como um importante elemento para a Web Semântica, devido à capacidade de programas conseguirem utilizar o XML para a resolução de uma série de tarefas. Devido a tais características, os dados em RDF passaram a ser majoritariamente publicados nesse formato, pelo menos nos primeiros quinze anos após a proposta da Web Semântica.

Gandon e Schreiber (2014) relatam que o RDF/XML segue as estruturas conceituais do RDF, convertendo os nós e os predicados nos termos do XML: nomes de elementos, nomes de atributos, conteúdo de elementos e valores de atributos. Os autores destacam ainda que o RDF/XML possibilita a definição dos *namespaces* por meio da estrutura XML-QNames, além de permitir inserir os URIs nos recursos que fazem parte da representação dos recursos do RDF, por meio de atributos do XML.

Ressalta-se também o modo como o RDF/XML diferencia a representação dos recursos e dos predicados, o que pode ser percebido na figura 6, que apresenta um exemplo de

RDF/XML. Salienta-se, ainda, que as linhas em vermelho (linhas 2, 4, 6 e 8) fazem referência a recursos, enquanto as linhas em amarelo (linhas 3 e 7) fazem referência a uma propriedade.

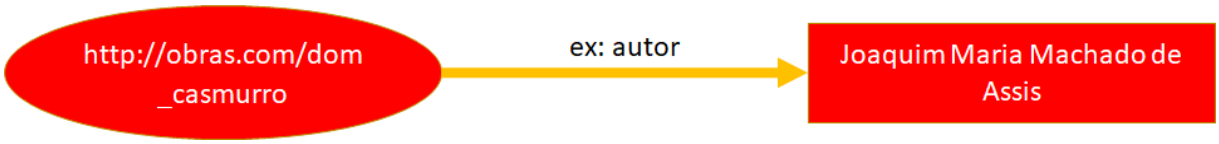
**Figura 6** - Exemplo de representação do RDF/XML

```
1. <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
2. <rdf:Description rdf:about="http://obras.com/dom_casmurro">
3.   <ex:autor>
4.    <rdf:Description rdf:about="http://autores.com/machado_de_assis">
5.          <nome>Joaquim Maria Machado de Assis</nome>
6.   </rdf:Description>
7.  </ex:autor>
8. </rdf:Description>
```

Fonte: elaborado pelo autor.

O código XML apresentado aponta o modo que o RDF pode ser descrito seguindo essa linguagem de marcação, demonstrando como as triplas RDF são concebidas nesse modelo. O código da figura 6 é embasado na estrutura conceitual apresentada pela figura 7.

**Figura 7** - Exemplo grafo RDF




Fonte: elaborado pelo autor.

A comparação entre as figuras 6 e 7 permite visualizar que os elementos foram mantidos, sendo, portanto, o XML uma opção de utilizar o modelo RDF mantendo as características conceituais e sendo um padrão utilizado em grande parte das aplicações computacionais. Há, porém, outros modelos que podem ser utilizados, como o N-Triple e o Turtle.

Carothers e Seaborne (2014) apontam que o N-Triple é: “A declaração de tripla [...] [com uma] sequência de termos (sujeito, predicado, objeto), separados por espaço em branco e terminados por '.' depois de cada tripla.”. Um exemplo de N-Triple pode ser visualizado na figura 8.

**Figura 8** - Exemplo N-Triple

```
<http://obras.com/dom_casmurro>        ex: autor        “Joaquim Maria Machado de Assis”.
```

Fonte: elaborado pelo autor.

O exemplo apresentado na figura 8 é embasado no grafo da figura 7. Verifica-se que essa estrutura é mais simples e fácil para compreender, tendo basicamente uma transposição do grafo para um formato textual.

Outro formato popular é o Turtle. Beckett et al. (2014) apresentam um documento com todas as especificações técnicas do Turtle. Os autores definem ainda o Turtle como uma representação textual dos grafos RDF. Visualize-se um exemplo na figura 9.

**Figura 9** - Exemplo Turtle

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://exemplo.net>

<http://obras.com/dom_casmurro>
        ex:autor        “Joaquim Maria Machado de Assis”.
```

Fonte: elaborado pelo autor.

Destaca-se que o formato Turtle tem várias semelhanças com o N-Triple, sendo facilmente reconhecido e identificável os elementos que fazem parte da tripla RDF. No entanto, ambos os formatos, apesar de serem de fácil leitura por pessoas, não são tão reconhecidos em aplicações computacionais, como o XML. Por isso, o XML acabou se popularizando mais, e esses dois formatos são utilizados com fins mais acadêmicos e em algumas poucas aplicações.

Mais recentemente, um outro formato começou a se destacar, o JSON-LD (JSON para *Linked Data)*. Isso está ocorrendo, pois o formato JSON passou a ser amplamente utilizado em aplicações computacionais, sendo em vários casos um substituto do XML para a troca e a interoperabilidade de dados. Segundo Crockford (2019), “JSON (*JavaScript Object Notation* - Notação de Objetos JavaScript) é uma formatação leve de troca de dados. Para seres humanos, é fácil de ler e escrever. Para máquinas, é fácil de interpretar e gerar.”

O JSON-LD é uma forma de publicar os dados para *Linked Data* em JSON. Na sua essência, o JSON-LD é uma forma de utilizar o RDF, tendo o JSON como estrutura. Sporny et al. (2019) apontam que:

> JSON-LD é uma sintaxe leve para serializar dados vinculados em JSON. Seu design permite que o JSON existente seja interpretado como Dados Vinculados com alterações mínimas. O JSON-LD destina-se principalmente a ser uma maneira de usar o *Linked Data* em ambientes de programação baseados na Web, para construir serviços da Web interoperáveis e para armazenar dados vinculados em mecanismos de armazenamento baseados em JSON.

As informações relatadas pelos autores demonstram que o princípio do JSON-LD está em ser um modo de estruturar informações, seguindo o modelo RDF, de acordo com a estrutura do JSON. Destaca-se que há um direcionamento para a questão da publicação de dados em *Linked Data*, que será explorado posteriormente, mas essa publicação de dados segue os princípios do RDF, como relatado por Sporny et al (2019), ao afirmarem que o "JSON é uma sintaxe concreta dos conceitos do RDF."

Sporny et al. (2019) ainda apontam algumas características do JSON-LD:

- Mecanismo para identificar unicamente em objetos JSON;
- Meio para desambiguar objetos compartilhados utilizando URIs;
- Meio para fazer referência a um site da Web por meio de um objeto JSON;
- Mecanismo para associar os diversos tipos de dados com valores como datas e horas;
- Mecanismo para facilmente expressar um ou mais grafos direcionados.

A figura 10 demonstra um exemplo de JSON-LD.

**Figura 10** - Exemplo JSON-LD

```
{
  "http://schema.org/name": "Manu Sporny",
  "http://schema.org/url": { "@id": "http://manu.sporny.org/" },  ← The '@id'
keyword means 'This value is an identifier that is an IRI'
  "http://schema.org/image": { "@id": "http://manu.sporny.org/images/manu.png" }

}
```

Fonte: Sporny et al. (2019).

A figura 10 demonstra como as triplas RDF são estruturadas nessa estrutura, mantendo os conceitos do modelo RDF. Vale destacar que diversos websites estão inserindo informações estruturadas em JSON-LD para auxiliar no processo de busca, retomando conceitos trazidos por Berners-Lee, Hendler e Lassila (2001), no texto inicial da Web Semântica.

A abrangência do formato JSON está levando os conceitos do RDF a várias áreas e em projetos de grande destaque. Em uma página de documentação, o Google recomenda o uso de

JSON-LD para que as páginas Web insiram dados estruturados para aprimorar as pesquisas realizadas. Nessa documentação, a empresa relata ainda que o JSON-LD:

> É uma notação JavaScript incorporada em uma tag <script> no cabeçalho ou no corpo da página. A marcação não é intercalada com texto visível para o usuário, o que facilita a expressão de itens de dados aninhados, como o país de um PostalAddress de um MusicVenue de um evento. Além disso, o Google pode ler dados JSON-LD quando eles são injetados dinamicamente no conteúdo da página, como por código JavaScript ou widgets incorporados no seu sistema de gerenciamento de conteúdo. (GOOGLE, 2019).

O uso e a recomendação de uma empresa como a Google, uma das grandes empresas da área de TI, faz com que a abrangência e a popularidade das ferramentas da Web Semântica aumentem bastante. Além disso, o uso do JSON é muito recomendado em aplicações computacionais, o que favorece a escolha pelo formato JSON-LD.

As quatro formas destacadas, RDF/XML, N-Triple, Turtle e JSON-LD, são a materialização do conceito do RDF, ao mesmo tempo que demonstram que as ferramentas da Web Semântica evoluíram significativamente nos últimos anos. Além disso, a proposta do JSON-LD indica que a Web Semântica permanece evoluindo e não está sendo utilizada apenas nos muros das universidades, pois quando criam uma materialização do conceito do RDF, tantos anos após a sua proposta, é visível a importância de se utilizar o RDF nas tecnologias mais atuais existentes.

Outra tecnologia fundamental para o presente trabalho é o OWL, linguagem para ontologias, que será detalhada a seguir.

### 3.3.2 OWL

Há diversas linguagens para a construção de ontologias, que possuem mais ou menos expressividade. Nesse contexto, como relatado anteriormente, devido a importância das ontologias para a Web Semântica, era necessário que fosse criada uma linguagem de ontologias capaz de efetivamente contextualizar e expressar significados, com um alto nível de semântica formal. Assim, a W3C propôs a *Web Ontology Language* (OWL).

Coneglian e Santarem Segundo (2018, p. 38-39) trazem essa questão, afirmando que:

> [...] o termo ontologia se refere a um conceito; assim, para existir uma ontologia, do ponto de vista prático e computacional, é necessário que seja utilizada alguma linguagem que implemente as características que envolvem as ontologias. A linguagem OWL foi desenvolvida com essa finalidade, buscando representar as características e propriedades, descrevendo computacionalmente os conceitos pertencentes às ontologias.

A afirmação dos autores traz o principal propósito da construção da linguagem OWL, que está vinculado à necessidade de implementar todos os conceitos que estão envolvidos em uma ontologia. Assim, a linguagem OWL foi iniciada a partir de uma revisão e complementação da linguagem de ontologia DAM+OIL, que não possuía as mesmas capacidades representacionais que o OWL.

A *World Wide Web Consortium* (2012a, tradução nossa) afirma que é

> [..] uma linguagem da Web Semântica projetada para representar o conhecimento rico e complexo sobre as coisas, grupos de coisas, e as relações entre as coisas. OWL é uma linguagem baseada em lógica computacional, tal que o conhecimento expresso em OWL pode ser explorado por programas de computador, por exemplo, para verificar a consistência de tal conhecimento ou para tornar o conhecimento implícito explícito. Documentos OWL, conhecidos como ontologias, podem ser publicados na *World Wide Web* e podem referir-se ou ser referidos de outras ontologias OWL. OWL faz parte da camada da Web Semântica do W3C, que inclui RDF, RDFS, SPARQL, etc.

A definição apresentada mostra alguns elementos que são centrais na linguagem OWL, como a questão da lógica computacional estar presente na ontologia, e conseguir ser compreendida pelo computador. Ademais, a possibilidade de relacionamentos com outras ontologias ou outras estruturas da Web Semântica, como RDF e SPARQL, fortalece o elemento da interoperabilidade nas ontologias. Ressalta-se que a linguagem OWL passa, assim, a ser a principal materialização de uma ontologia, pois cumpre o que as diversas definições traziam como elementos que são parte de uma ontologia, como uma especificação formal compartilhada.

Comprovando esses elementos, Santarem Segundo (2010, p. 128, grifo nosso) relata que: "A OWL foi projetada com o objetivo de ser **efetivamente utilizada** por aplicações que necessitem processar o conteúdo de informações, [...], a linguagem OWL é considerada **mais adaptada e mais fácil para expressar significados e semânticas** [...]". Isso demonstra que a linguagem OWL é a principal linguagem para atender as necessidades de expressividade semântica da Web Semântica, sendo, assim, a principal linguagem de construção de ontologias existente.

A linguagem OWL está atualmente na sua segunda versão, datada de 2012. No geral, a versão do OWL 2 é bastante semelhante à versão OWL 1, inserindo novas funcionalidades, que inseriram mais expressividade para a linguagem. Além disso, essa linguagem definiu uma nova sintaxe e inseriu novas funcionalidades e restrições. (WORLD WIDE CONSORTIUM, 2012b)

No que tange ao objetivo para se desenvolver uma ontologia utilizando a linguagem OWL, Santarem Segundo (2010, p. 128) relata que ao:

> [...] construir ontologias, [deve-se] explicitar fatos sobre um domínio, definir indivíduos que fazem parte de um domínio e afirmações sobre ele, definir classes e propriedades destas classes, especificar como derivar consequências lógicas (fatos não literalmente presentes na ontologia, mas resultantes de sua semântica) e racionalizar sobre ontologias e fatos.

A afirmação do autor explicita alguns dos principais elementos que definem e fazem parte de uma ontologia construída com a linguagem OWL. Breitman (2005) detalha melhor esses pontos, relatando que há seis elementos básicos na OWL, que são: *namespaces*, cabeçalhos, classes, indivíduos, propriedades, e restrições de classes e propriedades. Tais elementos serão descritos na sequência.

O primeiro elemento destacado é o *namespace*. O *namespace* trata de um conjunto de recursos que estão localizados em um mesmo domínio, o que é refletido no uso de URI com o mesmo início. Os arquivos de ontologias devem iniciar realizando a definição de uma série de *namespaces*, que serão utilizados para apontar interligações com outros recursos, favorecendo, assim, a interoperabilidade de recursos. Destaca-se, ainda, que a própria ontologia deve ter um *namespace* definido para os recursos dela. Um exemplo de *namespace* é o do próprio RDF, que deve ser definido para a utilização de todos os recursos, classes e elementos que estão vinculados àquele modelo.

A figura 11 aponta o início de uma declaração de ontologia, iniciando, portanto, por um conjunto de *namespaces*.

**Figura 11** - Exemplo de definição de *namespace*

```
<rdf:RDF
  xmlns     =     "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
  xmlns:vin =    "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
  xml:base  =    "http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#"
  xmlns:food=   "http://www.w3.org/TR/2003/PR-owl-guide-20031209/food#"
  xmlns:owl =   "http://www.w3.org/2002/07/owl#"
  xmlns:rdf =    "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs=  "http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd =    "http://www.w3.org/2001/XMLSchema#">
```

Fonte: elaborado pelo autor.

A figura 11 insere um conjunto de *namespaces*, que serão utilizados na ontologia para fazer referência a objetos que já foram definidos em outros domínios, ou serão criados naquela ontologia.

Após a inserção dos *namespaces*, as ontologias contêm os cabeçalhos. O cabeçalho apresenta algumas informações que fazem referência a própria ontologia, com informações sobre datas, versões, além de inserir conceitos e propriedades que são de outras ontologias.

O próximo elemento a ser considerado é a classe, que é a principal estrutura para definição dos conceitos e dos termos que fazem parte da ontologia, sendo esse elemento que irá representar algo do mundo real na ontologia. Breitman (2005, p. 61) afirma que "[...] uma classe representa um conjunto ou coleção de indivíduos (objetos, pessoas, coisas) que compartilham de um grupo de características que os distinguem dos demais. Utilizamos classes para descrever conceitos de um domínio, por exemplo, móveis [...]".

A classe é, assim, uma forma de abstrair um conceito, representando uma série de elementos que são classificados dentro desse elemento. Nesse sentido, Santarem Segundo (2010, p. 133) afirma que as classes:

> [...] são responsáveis por representar um grupo de indivíduos com características comuns, provendo um mecanismo de abstração para agrupar recursos com características similares, ou seja, as classes têm a característica de representar um conjunto ou uma coleção de indivíduos que compartilham das mesmas características.

Um outro aspecto acerca das classes do OWL é trazido por Bechhofer et al. (2004, tradução nossa):

> Dois identificadores de classe OWL são predefinidos: as classes owl:Thing e owl:Nothing. A extensão de classe owl:Thing é o conjunto de todos os indivíduos. A extensão de classe da owl:Nothing é o conjunto vazio. Conseqüentemente, toda classe OWL é uma subclasse de owl:Thing e owl:Nothing é uma subclasse de toda classe.

O conceito de que toda classe é uma subclasse de owl:Thing é importante para entender que todas as classes estão, de algum modo, interligadas, e herdam essas características. Essa relação possibilita a compreensão de que toda classe é uma abstração de algo, que herda características de alguma outra classe. Além disso, a ontologia em OWL possibilita a definição de subclasses, que herdam as características da classe mãe, como na interpretação dada para as classes no paradigma de programação de orientação a objetos.

A figura 12 apresenta um exemplo de como é definida uma classe, além de apresentar como uma classe é denominada como subclasse de outra.

**Figura 12** - Declaração de classe e subclasse

```
<owl:Class rdf:ID="Música"/>

<owl:Class rdf:ID="Opera">
  <rdfs:subClassOf rdf:resource="#Musical" />
</owl:Class>
```

Fonte: elaborada pelo autor.

Na figura 12, a primeira linha apresenta a definição da classe “Música”, enquanto as linhas seguintes apresentam a definição da classe “Opera”, como uma subclasse da classe “Musical”. Vale destacar que a definição de uma subclasse ocorre por meio da propriedade “rdfs:subClassOf”, que é oriunda do vocabulário do RDF Schema.

O próximo elemento que faz parte das ontologias é o indivíduo. O indivíduo é uma concretização de um conceito abstrato que é uma classe, sendo, portanto, algum objeto do mundo real, estando assim, sempre vinculado a alguma classe definida. Bechhofer et al. (2004, tradução nossa) complementam essa questão, relatando que “[...] os indivíduos são instâncias de classes, e as propriedades podem ser usadas para relacionar um indivíduo a outro.”

A figura 13 aponta como pode ser declarado um indivíduo.

**Figura 13** - Declaração de indivíduo

```
<Opera rdf:ID="Tosca">
   …
</Opera>
```

Fonte: elaborada pelo autor.

A declaração realizada na figura 13 aponta uma instância (indivíduo) chamado “Tosca” da classe “Opera”. Ou seja, “Tosca” é um objeto real da abstração “Opera”; sendo assim, uma representação dessa classe, com todas as suas características.

O elemento seguinte é a propriedade, que insere características e relacionamentos para as classes. Santarem Segundo (2010, p. 135) afirma que “[...] propriedades são recursos da linguagem OWL que têm o propósito de descrever fatos em geral. As propriedades são utilizadas para estabelecer relacionamentos entre os indivíduos ou ainda entre indivíduos e valores.”

As propriedades estão divididas em dois tipos, as propriedades de dados e as propriedades de objetos.

As propriedades de dados (*data properties*) são aquelas que vinculam um objeto a um dado propriamente, e são sempre informações literais, como valores inteiros, *strings* ou valores booleanos. Um exemplo de uma propriedade de dados é a propriedade "nome".

As propriedades de objetos (*object properties*) são aquelas que relacionam dois objetos distintos, e possibilitam que as relações existentes entre os dados sejam mais expressivas e com um maior nível de semântica formal. Um exemplo de uma propriedade de objeto seria a classe "CD" estar ligado a classe "Música" por meio de uma propriedade de objeto "tem_faixa".

Por fim, o último elemento de uma ontologia OWL é a restrição. A restrição é um dos principais elementos para fornecer um alto nível de expressividade e capacidade de formalizar as regras de um domínio, característica fundamental em uma ontologia. As restrições podem ter duas funções: restringir os valores que podem ser inseridos em uma propriedade ou apontar a cardinalidade com que dois objetos podem estar vinculados.

Coneglian (2017, p. 53, grifo nosso) aponta os tipos de relações de cardinalidade existentes:

> Existem três tipos de restrições de cardinalidade: "**owl:maxCardinality**", que descreve o número máximo de elementos distintos que uma classe deve possuir em uma determinada propriedade; "**owl:minCardinality**", contrária à anterior, definindo o número mínimo de elementos distintos que uma classe deve possuir em uma propriedade, e "**owl:cardinality**", aponta o número exato de elementos distintos que a classe deve possuir na propriedade. A segunda versão da OWL [...] insere ainda a possibilidade de inserção de cardinalidade qualificada, por meio das propriedades "**owl:maxQualifiedCardinality**", "**owl:minQualifiedCardinality**" e "**owl:qualifiedCardinality**", que apresentam as mesmas características que "owl:maxCardinality", "owl:minCardinality" e "owl:cardinality", respectivamente.

No que diz respeito às restrições de valores, tem-se quatro tipos de restrições: "*owl:allValuesFrom", "owl:someValuesFrom", "owl:hasValue" e "owl:hasSelf*", assim caracterizadas por Coneglian (2017, p. 53): o *owl:allValuesFrom*: "[...] tem a função de definir quais são os valores possíveis que uma propriedade determinada pode ter.", o *owl:someValuesFrom*: "[...] tem a função de determinar a classe e a ocorrência de pelo menos um valor dentre as propriedades" e a propriedade *owl:hasValue* "[...] tem a função de definir um valor determinado que uma determinada propriedade especificada pode possuir." Por fim, sobre a propriedade *owl:hasSelf,* o autor relata que: "[...] tem a função de especificar uma ligação de um indivíduo nele mesmo."

Um terceiro tipo de restrição é chamado de propriedade especial. Essas propriedades são aquelas responsáveis por definir os axiomas e a realização de inferências, visando a inserir uma carga semântica mais elaborada e forte nas ontologias construídas em OWL. Coneglian (2017) realiza um levantamento e um aprofundamento de todas as propriedades, o que não será realizado aqui, pois não será utilizada propriedade por propriedade, mas sim, uma visão mais geral desse tipo de restrição.

Nesta subseção foi apresentada uma visão mais geral do OWL, e quais são os elementos que são parte dessa linguagem. A linguagem OWL será essencial para a conceituação deste trabalho, pois define quais são os elementos de uma ontologia que podem ser utilizados.

A seguir apresentam-se outras ferramentas da Web Semântica que têm destaque e são necessárias para o desenvolvimento deste trabalho.

### 3.3.3 Outras ferramentas (SKOS, SWRL, SPARQL)

Enfatizam-se aqui outras ferramentas da Web Semântica, fundamentais para o desenvolvimento deste trabalho.

A primeira ferramenta é o *Simple Knowledge Organization System* (SKOS), que é definido como: "[...] uma área de trabalho que desenvolve especificações e padrões para apoiar o uso de sistemas de organização do conhecimento (KOS), como tesauros, esquemas de classificação, listas de cabeçalhos de assuntos e taxonomias dentro da estrutura da Web Semântica." (WORLD WIDE WEB CONSORTIUM, 2012c, tradução nossa).

O SKOS foi criado como uma ferramenta, aderente aos conceitos da Web Semântica para a criação de sistemas de organização do conhecimento, em especial de tesauros. O SKOS fornece uma estrutura para a criação desses instrumentos, ao mesmo tempo que possui uma série de termos que permitem a criação das relações que os tesauros possuem.

Santarem Segundo e Coneglian (2015, p. 227) afirmam que:

> O modelo SKOS divide-se em 8 grupos: conceitos de classe, conceitos de esquema, rótulos, notações, propriedades, relações semânticas, coleções e mapeamento de propriedades. Esses grupos são representados por elementos (como: skos:prefLabel, skos:altLabel, entre outros) nomeados vocabulários, que efetivam a relação existente entre os termos em um vocabulário controlado.

Verifica-se pelas afirmações dos autores que há diversas semelhanças entre o SKOS e os elementos das ontologias OWL, destacando-se um foco grande nos vocabulários e no compartilhamento. Isso pode ser visualizado na afirmação de Miles e Bechhofer (2009, tradução nossa), no documento que é a referência técnica do SKOS: "[...] [O SKOS] é um

padrão de compartilhamento de dados, conectando vários campos diferentes de conhecimento, tecnologia e prática.”

O SKOS está vinculado tanto ao RDF quanto ao OWL. A relação com o primeiro está no modo como o SKOS é organizado e estruturado; assim, os dados do SKOS são expostos em triplas RDF. Quanto ao OWL, o SKOS pode ser utilizado junto com ele, para expressar e trocar informações sobre um domínio.

A figura 14 mostra um exemplo de dados utilizando o SKOS.

**Figura 14 -** Exemplo de dados SKOS

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

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

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

Fonte: Miles e Bechhofer (2009).

O trecho apresentado na figura 14 demonstra que a estrutura seguida é a utilizada no RDF, inserindo uma série de elementos que são relativos aos tesauros e que são inseridos com o prefixo “skos”.

Outra tecnologia de importância para o presente trabalho é *Semantic Web Rule Language* (SWRL). A linguagem SWRL é uma combinação da linguagem OWL com sublinguagens do *Rule Markup Language*. O princípio dessa linguagem está em estender os axiomas do OWL, adicionando outros tipos de regras.

Horrocks et al. (2004, tradução nossa) afirmam que:

> As regras propostas [do SWRL] são da forma de uma implicação entre um antecedente (corpo) e consequente (cabeça). O significado pretendido pode ser lido como: sempre que as condições especificadas no antecedente são mantidas, as condições especificadas no consequente também devem ser mantidas.

Em suma, a lógica apontada na linguagem SWRL aponta que caso a expressão seja verdadeira, deverá ser realizada uma ação de acordo com o expresso no axioma. Isso é

interessante, pois aumenta as possibilidades de inferências e axiomas que podem ser realizados com o OWL. A ideia é ser uma linguagem complementar às ontologias, que permita a inserção de mais axiomas lógicos e, consequentemente, permite mais inferências.

Santarem Segundo e Coneglian (2016, p. 226) apontam que: “Em síntese, a criação de regras construídas com SWRL contém duas partes essenciais; a primeira parte, um conjunto de condições, e a segunda parte que contém a inferência a ser executada caso a primeira parte da expressão seja verdadeira.”

A figura 15 mostra um exemplo de informações e inferências inseridas utilizando o SWRL.

**Figura 15** - Exemplo de inferências em SWRL

```
<swrlx:classAtom>
  <owlx:Class owlx:name="Person" />
  <ruleml:var>x1</ruleml:var>
</swrlx:classAtom>

<swrlx:classAtom>
  <owlx:IntersectionOf>
    <owlx:Class owlx:name="Person" />
    <owlx:ObjectRestriction owlx:property="hasParent">
      <owlx:someValuesFrom owlx:class="Physician" />
    </owlx:ObjectRestriction>
  </owlx:IntersectionOf>
  <ruleml:var>x2</ruleml:var>
        </swrlx:classAtom>
```

Fonte: Horrocks et al. (2004).

Na figura 15 é possível visualizar algumas informações e relações realizadas, demonstrando como o axioma é construído, utilizando as *tags* e as estruturas do XML e do RDF.

A última tecnologia a ser apresentada aqui será o *SPARQL Protocol and RDF Query Language* (SPARQL). Esse é o protocolo recomendado pela W3C para a realização de consultas e buscas, e utiliza os princípios do RDF para recuperar as informações em bases de dados estruturadas de RDF e OWL.

O SPARQL é bastante importante para a Web Semântica, pois o uso efetivo de todos os conceitos e princípios, especialmente do RDF, só é possível em sua totalidade por meio do SPARQL. Neste âmbito, Berners-Lee afirmou que “[...] tentar usar a Web Semântica sem SPARQL é como tentar usar um banco de dados relacional sem SQL. SPARQL torna possível

consultar informações de bancos de dados e outras fontes diversas em estado natural, em toda a Web." (WORLD WIDE WEB CONSORTIUM, 2007, tradução nossa).

Sobre o funcionamento do SPARQL, Santarem Segundo (2014, p. 3870) relata que esse protocolo: "[...] é um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web." A afirmação do autor, juntamente com o relatado por Berners-Lee, demonstra que o SPARQL é necessário para a obtenção e a utilização de dados que seguem os princípios da Web Semântica. Nesse contexto, Coneglian (2017, p. 57) afirma que: "[...] uma extração utilizando tecnologias tradicionais de banco de dados, como SQL, [em dados seguindo os princípios da Web Semântica] não permitiria uma obtenção de todas as características que tais representações possuem."

A figura 16 demonstra um exemplo de código em SPARQL.

**Figura 16** - Exemplo de código em SPARQL

```
PREFIX p: <http://tese.com/ppgci/pessoa#>
PREFIX b: <http://tese.com/ppgci/banco#>
SELECT ?id ?nome ?saldo
WHERE
{
?pessoa     p:nome      ?nome ;
            p:id        ?id .
?banco      p:id        ?id ;
            b:saldo     ?saldo .
}
```

Fonte: elaborado pelo autor.

O exemplo apresentado na figura 16 busca informações sobre saldo de uma pessoa, relacionando essas informações na estrutura base do RDF. Destaca-se que, apesar de algumas semelhanças de sintaxe com o SQL, o SPARQL tem uma lógica distinta, em que as relações devem acontecer a partir das triplas RDF.

O SPARQL, por tais características, é capaz de recuperar informações em bases de RDF e OWL, utilizando e explorando toda a expressividade semântica desses modelos, sendo, portanto, fortemente recomendada a sua utilização, quando se trata de dados no contexto da Web Semântica.

Todos os conceitos e as ferramentas da Web Semântica apontadas nas últimas subseções foram e continuam sendo essenciais para o processo chamado de materialização da Web Semântica, que será melhor explorado na próxima subseção, juntamente com a explanação acerca do *Linked Data*.

## 3.4 MATERIALIZAÇÃO DA WEB SEMÂNTICA E *LINKED DATA*

A Web Semântica, desde o seu início, em 2001, está passando por um processo de maturação, em que essa proposta deixou de ser apenas teórica e conceitual, e passou a ser aplicada e prática, havendo iniciativas reais e implementadas demonstrando a sua eficiência.

Assim, diversos autores vêm afirmando que esse processo deve ser chamado de materialização da Web Semântica, pois, a partir do momento que a Web Semântica começa a ter aplicações desenvolvidas em seu contexto, ela deixa de ser apenas teórica, e passa a ser real e aplicada. Santarem Segundo (2014, p. 3864) considera que as tecnologias e as ferramentas da Web Semântica, entre elas, RDF, OWL, XML, bem como os seus conceitos: "[...] tornam possível a materialização do conceito da Web Semântica."

Posteriormente, Coneglian e Santarem Segundo (2018) apontam que a materialização da Web Semântica está na definição de aplicações e no uso das tecnologias e das ferramentas da Web Semântica nos diversos contextos. Além disso, todos esses autores apontam para o fato de que a principal materialização da Web Semântica está na proposta do *Linked Data*, pois ela utiliza conceitos, tecnologias e ferramentas da Web Semântica em uma aplicação prática e que está sendo utilizada significativamente.

Essa discussão ocorre, porque a Web Semântica, no seu início, estava mais direcionada para estudos teóricos, com discussões sobre as suas camadas, e algumas tecnologias, sem que houvesse grupos e trabalhos que visassem, de fato, a tornar a proposta real. Dessa forma, demorou alguns anos até que a Web Semântica, junto às suas ferramentas e conceitos, tivesse aplicações implementando todos os princípios e os benefícios existentes.

A criação do *Linked Data* começou a alterar esse cenário, mas junto a sua criação, diversas outras propostas começaram a ser realizadas, como o uso das ferramentas da Web Semântica em buscadores e na recuperação da informação, além do uso em redes sociais e outros.

O quadro 4 sintetiza essa discussão, ao pontuar como os conceitos, as ferramentas e as aplicações da Web Semântica contribuíram para o processo de sua materialização.

**Quadro 4** - Materialização da Web Semântica

| Elementos | Descrição | Nível de materialização |
|---|---|---|
| Conceitos da Web Semântica | Processo de definição do que é a Web Semântica e conceitos-chave que foram importantes. Exemplos como ontologias e o conceito RDF. | Materialização baixa, apesar de essas serem as bases para todas as etapas seguintes. |

| Elementos | Descrição | Nível de materialização |
|---|---|---|
| Ferramentas da Web Semântica | Inicia junto com o processo de conceituação, mas foi um processo mais lento, devido à necessidade de que houvesse conceitos mais claros. Posteriormente, o processo de desenvolvimento de ferramentas ocorreu muito rapidamente, havendo a definição e criação da maioria das ferramentas em um curto período. | Inicia-se o processo de materialização, pois, as ferramentas criadas começam a demonstrar como a Web Semântica poderia ser utilizada na prática. No entanto, no início, não havia demonstrações de aplicações de tais ferramentas. |
| Aplicações da Web Semântica | Com a criação de ferramentas, e havendo conceitos bem claros da Web Semântica, iniciou-se o processo de criação e desenvolvimento de aplicações capazes de mostrar o potencial e a importância da Web Semântica. O *Linked Data* é considerado o principal elemento dessa materialização (CONEGLIAN E SANTAREM SEGUNDO, 2018). | Materialização alta, pois, com essas aplicações verifica-se que a Web Semântica se torna real e implementável, incentivando o desenvolvimento de uma série de outras aplicações. |

Fonte: elaborado pelo autor.

O quadro 4 demonstra como a materialização da Web Semântica foi sendo alcançada a partir dos elementos, conceitos, ferramentas e aplicações. Dessa forma, visualiza-se que no momento atual, a Web Semântica está madura e podendo avançar e contribuir para outras áreas do conhecimento, como na recuperação da informação e na Inteligência Artificial, pontos que serão utilizados e discutidos neste trabalho.

Como relatado no quadro, o principal elemento dessa materialização da Web Semântica é o *Linked Data*, que será explorado a seguir.

Berners-Lee, em 2006, apresentou a proposta do *Linked Data*, buscando melhorar e aprimorar a própria proposta da Web Semântica. Berners-Lee (2006, tradução nossa), nesse texto, relatou: “A Web Semântica não é apenas sobre colocar dados na Web. Trata-se de criar links, para que uma pessoa ou uma máquina possa explorar a rede de dados. Com o *Linked Data*, quando você tem alguns dados, pode encontrar outros dados relacionados.”

A afirmação do autor demonstra a necessidade de explorar e de existir redes de dados ligados que possibilitem o encontro e a exploração de tais dados. Além disso, Berners-Lee posiciona a relação entre Web Semântica e *Linked Data* por meio desse trecho.

No contexto do *Linked Data*, vale ainda destacar a seguinte questão:

> A web semântica permite que as máquinas obtenham significado a partir de dados estruturados que podem ser processados quando publicados como dados vinculados. Os dados vinculados são uma etapa fundamental da implementação em direção à web semântica, na qual representar entidades de informações por meio de URIs as torna processáveis por máquina. (MÉNDEZ; GREENBERG, 2012, p. 238, tradução nossa)

Os autores demonstram que o *Linked Data*, tratado como dados vinculados, é essencial para que a Web Semântica possa processar as informações e, efetivamente, consiga obter o significado das informações. Dessa forma, um dos pontos principais do *Linked Data* é apresentado, tratando da questão do uso de URIs para os recursos.

A questão da utilização de URIs pelos recursos foi apresentada na proposta inicial do *Linked Data*, realizado por Berners-Lee (2006, tradução nossa), que assim apresenta os quatro princípios dessa proposta:

> 1. Usar URIs como nomes para as coisas; 2. Usar HTTP URIs para que as pessoas possam procurar esses nomes; 3. Quando uma pessoa procura uma URI, fornecer informações úteis, usando padrões (RDF *,SPARQL); 4. Incluir links para outras URIs para que ela possa descobrir mais coisas.

O primeiro princípio (“Usar URIs como nomes para as coisas”) trata da importância da URI, devido à necessidade de haver identificadores únicos nos recursos, evitando, assim, a ambiguidade dos dados, o que torna a URI essencial para que as informações possam ser relacionadas de forma correta. O segundo princípio (“Usar HTTP URIs para que as pessoas possam procurar esses nomes”) visa a melhorar o encontro e a recuperação da informação, de modo que, quando uma pessoa se deparar com uma URI, ela será capaz de encontrar essa informação na Web. O terceiro princípio (“Quando alguém procura uma URI, fornecer informações úteis, usando padrões (RDF *,SPARQL)”) demonstra a necessidade de se utilizar os ferramentais da Web Semântica para que se possam publicar os dados, o RDF, e para se recuperar os dados, o SPARQL.

Uma outra definição do *Linked Data* é dada por Bizer, Heath e Berners-Lee (2009, tradução nossa): “Tecnicamente, *Linked Data* refere-se a dados publicados na Web de forma que sejam legíveis por máquina, seu significado é definido explicitamente, está vinculado a outros conjuntos de dados externos e pode, por sua vez, ser vinculado por conjuntos de dados externos.”

A presente definição traz alguns importantes elementos que pautaram o modo como a proposta do *Linked Data* cresceu e se desenvolveu, como a necessidade de conjuntos de dados estarem publicados na Web, estar vinculado a outros conjuntos de dados e ser referenciado por

outros conjuntos de dados. A ideia está em criar uma grande rede de dados, que possui ligações em diversas direções.

Nesse sentido, Heath e Bizer (2011, tradução nossa) reafirma que:

> O termo *Linked Data* refere-se a um conjunto de melhores práticas para a publicação e interligação de dados estruturados na Web. Essas boas práticas foram introduzidas por Tim Berners-Lee em sua arquitetura de dados denominada *Linked Data* e tornaram-se conhecidas como os princípios de *Linked Data*.

A afirmação dos autores demonstra que o termo *Linked Data* diz respeito às regras e à maneira como os dados devem ser publicados; a partir desta proposta, iniciou-se um processo de publicação de vários conjuntos de dados, seguindo as regras do *Linked Data* e os princípios da Web Semântica. Adicionalmente, esses *datasets* começaram a utilizar diversos vocabulários, visando ao favorecimento da interoperabilidade, ao fornecer o uso de padrões comuns e compartilhados.

Dentro desse cenário, surgiu a iniciativa do *Linking Open Data* (LOD), que começou a reunir um grande conjunto de dados publicados seguindo os padrões do *Linked Data.* A ideia foi criar uma nuvem de dados abertos, publicados seguindo os princípios do *Linked Data*.

Santarem Segundo (2015, p. 225) afirma sobre esse projeto que:

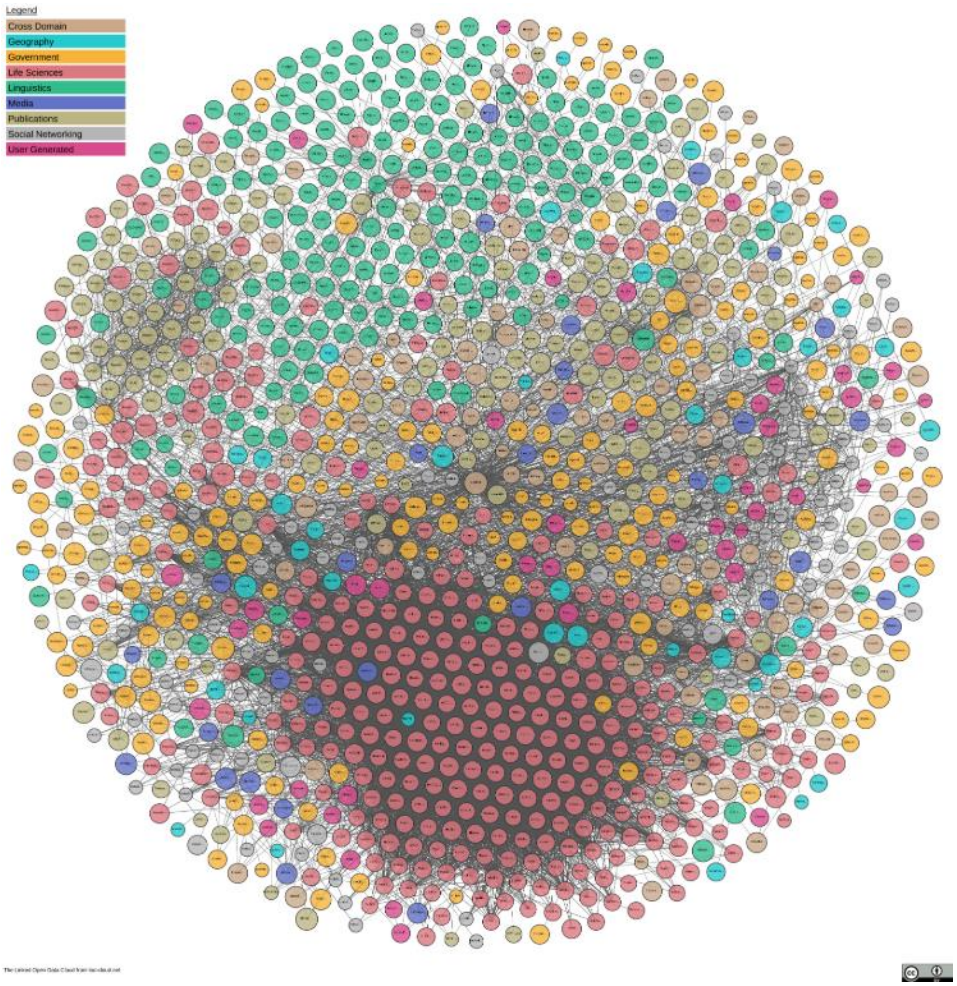
> O LOD, que atualmente apresenta-se como a melhor forma de materialização dos conceitos e tecnologias da Web Semântica, é um projeto, com um conjunto de normas a serem seguidas, que usa os mesmos princípios de ligação semântica da Web de Dados. Entretanto tem particularidades específicas, indicando um grau de exigência maior na constituição de sua rede de interligações.

Como relatado pelo autor, a iniciativa do LOD possui algumas exigências para a inserção de seus dados nessa nuvem. As exigências vão desde a necessidade de utilizar vocabulários reconhecidos nos dados, até a necessidade de fazer referências aos *datasets* que já fazem parte da iniciativa.

Atualmente, o projeto é chamado de *The Linked Open Data Cloud* e no mês de março de 2019, possuía 1.239 conjuntos de dados e 16.147 ligações, contendo conjuntos de diversas áreas como: domínio geral, geografia, governo, ciências da saúde, linguística, mídia, publicações, redes sociais e geradas por usuários.

A figura 17 mostra a nuvem dos conjuntos de dados atualizados na data de março de 2019.

**Figura 17** - Conjunto de dados do *The Linked Open Data Cloud*




Fonte: *The Linked Open Data Cloud* (2019).

Por meio da figura 17, é possível visualizar a quantidade de relações e como os *datasets* estão vinculados. Isso indica que muitos conjuntos de dados utilizam e fazem referências a outros *datasets*, promovendo, assim, o que a proposta do *Linked Data* inicialmente desejava.

Vale destacar que todos esses conjuntos de dados utilizam as ferramentas da Web Semântica, como o RDF e o OWL, o que torna os dados ali contidos, com uma alta expressividade. Assim, ao usar linguagem de recuperação como o SPARQL, os conjuntos de dados do LOD são capazes de fornecer uma contextualização bastante elevada das informações, principalmente, pois, ao utilizar vocabulários e ontologias, a possibilidade de realização de inferências para a descoberta de significados aumenta significativamente.

Um último ponto importante para a questão da efetivação do *Linked Data* está no modo como os dados são publicados. Nesse sentido, há algumas ferramentas que auxiliam e colaboram para transformar os dados nos formatos do *Linked Data*, além de plataformas que

possibilitam disponibilizar os dados abertamente. Isotani e Bittencourt (2015) destacam uma série de tecnologias, como o CKAN, Virtuoso, Jena e Sesame, que permitem a publicação dos dados abertamente, possibilitando que eles sejam inseridos na nuvem do LOD.

A seguir apresentam-se os conceitos da área de Inteligência Artificial e processamento de linguagem natural, um campo vinculado, mas com distintas características no modo como a Web Semântica trata a questão da semântica formal.

# 4 INTELIGÊNCIA ARTIFICIAL

Esta seção irá destacar o terceiro conceito-chave deste trabalho, que é a Inteligência Artificial. A aproximação da área da Ciência da Informação com a Inteligência Artificial iniciou há vários anos, mas a própria evolução desta segunda possibilitou que, atualmente, essa aproximação possa ser rediscutida, se tornando mais forte e efetiva.

Este trabalho traz uma nova aproximação nesse sentido, trazendo campos de estudos e perspectivas da Ciência da Informação, aproximando-se da área da Inteligência Artificial e, em especial, da área do processamento de linguagem natural.

## 4.1 INTRODUÇÃO À INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial vem sendo foco de estudos de diversas áreas do conhecimento nos últimos anos. Esse campo de estudos que nasce vinculado à área de Ciência da Computação teve uma grande evolução, em especial pelo avanço tecnológico das últimas décadas, o que levou à criação de diversas tecnologias para o dia a dia das pessoas, tecnologias que usam a Inteligência Artificial e são nela embasadas.

A Inteligência Artificial é definida em quatro categorias principais: sistemas que pensam como humanos; sistemas que agem como humanos, sistemas que pensam de forma racional e sistemas que agem de forma racional. Verifica-se que a abordagem centrada nos humanos possui um caráter mais empírico, enquanto que a abordagem racional está mais focada nas questões matemáticas e ligadas à engenharia. (RUSSELL; NORVIG, 2016).

As quatro definições podem ser visualizadas no quadro 5, que representa cada classificação dada a cada definição.

**Quadro 5** – Definições de Inteligência Artificial

| **Sistemas que pensam como humanos** | **Sistemas que pensam racionalmente** |
|---|---|
| "O novo e empolgante esforço para fazer os computadores pensarem... Máquinas com mentes, no sentido pleno e literal" (HAUGELAND, 1985, tradução nossa)<br><br>"[A automação de] atividades que associamos ao pensamento humano, atividades como a tomada de decisão, resolução de problemas, aprendizado [...]" (BELLMAN, 1978, tradução nossa) | "O estudo das faculdades mentais através do uso de modelos computacionais" (CHARNIAK; MCDERMOTT, 1985, tradução nossa)<br><br>"O estudo dos cálculos que possibilitam perceber, raciocinar e agir" (WINSTON, 1992, tradução nossa) |
| **Sistemas que agem como humanos** | **Sistemas que agem racionalmente** |
| "A arte de criar máquinas que executam funções que requerem inteligência quando executadas por pessoas" (KURZWEIL, 1990, tradução nossa)<br><br>"O estudo de como fazer os computadores executarem coisas nas quais, no momento, as pessoas são melhores" (RICH; KNIGHT, 1991, tradução nossa) | "Um campo de estudo que procura explicar e imitar o comportamento inteligente em termos de processos computacionais" (SCHALKOFF, 1990, tradução nossa)<br><br>"O ramo da ciência da computação que se preocupa com a automação do comportamento inteligente" (LUGER; STUBBLEFIELD, 1993, tradução nossa) |

Fonte: Adaptado de Russell e Norvig (2016)

As distintas definições apresentadas no quadro 5 demonstram como a Inteligência Artificial foi concebida e pensada de diferentes modos, com alguns pensadores tendo uma perspectiva mais filosófica, em que as máquinas de fato se aproximam do comportamento humano, até um modo de ver mais pragmático, com um enfoque mais na matemática e na resolução de problemas. Essas diferentes visões influenciam a área da Inteligência Artificial até os dias atuais, com pesquisas que buscam criar serviços e produtos em cada uma das categorias citadas.

Um exemplo dessa evolução a partir de tais definições está em soluções criadas, como, por exemplo, os *chatbots*, inteligência artificial que busca agir com um humano, por meio de conversas, e sistemas de reconhecimento de placas, que pensam e agem de forma racional, uma vez que o objetivo é apenas identificar quais são as placas dos veículos de forma automatizada. Ambos os exemplos são Inteligências Artificiais que têm objetivos e concepções totalmente distintas.

A partir dessas abordagens, a área da Inteligência Artificial foi evoluindo de forma significativa, desde uma abordagem mais filosófica, que tratava de questões de cunho mais

teórico, passando pela abordagem matemática, com o estudo de algoritmos e probabilidades, pela abordagem psicológica, com um estudo mais cognitivo, chegando até às abordagens computacionais, com enfoque nos artefatos que podem ser construídos, e linguística, focada nas representações do conhecimento.

Todo esse processo, atualmente, permitiu que, ao tratar de Inteligência Artificial, se possa falar da existência de diferentes subcampos e áreas, que se propõem a resolver problemas dos mais diversos. Ainda segundo Russell e Norvig (2016), uma abordagem que tem um grande destaque atualmente e é vista com sucesso está ligada ao uso das redes neurais, que simulam o funcionamento do cérebro humano, para pensar em como ocorre o aprendizado das máquinas.

Hinton, Vinyals e Dean (2015) apontam que as redes neurais apresentam a probabilidade de uma classe ser parte de um grupo ou de uma determinada classificação. Esse processo acontece por meio da aplicação de um número de camadas, em que a partir dos valores da entrada, tem-se uma determinada saída.

As redes neurais são utilizadas em diversos campos da Inteligência Artificial, pois utilizando uma abordagem probabilística, é capaz de encontrar padrões e correlações existentes entre os dados. Por exemplo, os algoritmos de *machine learning* utilizam os princípios das redes neurais, justamente pelos resultados eficientes para realizar o treinamento e a identificação dos padrões existentes.

Outro importante elemento da Inteligência Artificial, que está diretamente vinculado à proposta da Web Semântica 4.0 são os agentes inteligentes. Nesse sentido, um agente é uma representação de algo que pode perceber o ambiente por meio de sensores e atuar nesse mesmo ambiente.

Nesse contexto, os agentes inteligentes utilizam elementos computacionais para perceberem, tais como câmeras, sensores e extratores de documentos, e para atuarem no mundo, tais como sistemas que enviam alguma informação, criação de relatórios e motores, que dependem de qual é o contexto de cada agente. (RUSSELL; NORVIG, 2016)

Em linhas gerais, um agente inteligente deve ser capaz agir racionalmente, o que pode ser avaliado a partir de métricas diversas. Além disso, a ideia de agente inteligente está na possibilidade dele poder ser treinado e ir se aperfeiçoando com o tempo, o que pode inclusive estar ligado à necessidade de ser corrigido constantemente por uma pessoa ou por outra máquina.

Outra importante área de estudos da Inteligência Artificial é o processamento de linguagem natural. Liddy (2001) aponta que esse campo de estudo utiliza abordagens simbólicas e estatísticas para permitir uma compreensão da fala e dos textos, de modo que, ao

criar sistemas, possa existir uma maior aproximação entre a linguagem computacional e a linguagem natural.

Todos esses campos estão contribuindo para tornar a Inteligência Artificial parte do cotidiano das pessoas. Ambientes criados mais recentemente, com ferramentas, bibliotecas e serviços, como o IBM Watson, AWS e Google Cloud, além de um grande conjunto de bibliotecas de programação nas mais diversas linguagens, tornaram a possibilidade de uso de técnicas de Inteligência Artificial algo mais simples, cuja solução pode ser integrada facilmente.

Destaca-se ainda que o uso de *chatbots* ou de análise de dados para a tomada de decisão, utilizam Inteligência Artificial em todo o processo e estão revolucionando o modo como as empresas vendem ou decidem o futuro de empregos, produtos e serviços. Esse cenário demonstra como a popularização dessas técnicas aconteceu de forma rápida e sustentável.

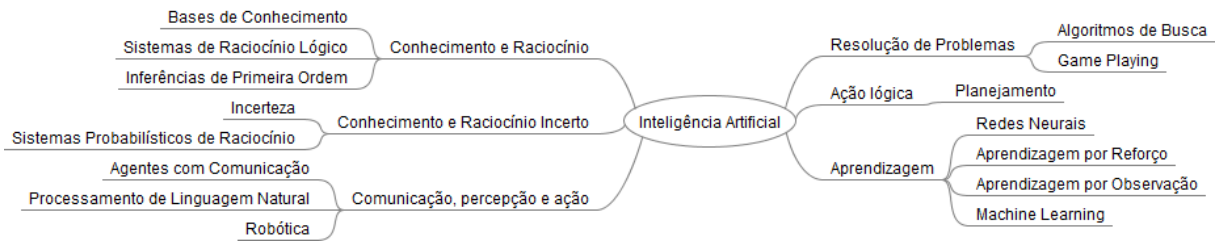
Aponta-se que existem diversos campos de estudos dentro da Inteligência Artificial, como processamento de linguagem natural, *machine learning*, *deep learning*, agentes inteligentes, entre outros.

## 4.2 CAMPOS DA INTELIGÊNCIA ARTIFICIAL

Há diversas áreas e subcampos da Inteligência Artificial que buscam atender propósitos e objetivos diferentes. Cada uma dessas áreas se destaca por algoritmos e técnicas diferentes, contribuindo para a evolução da área de formas distintas.

Destaca-se que existem diferentes classificações das áreas de Inteligência Artificial, com algumas tendo uma perspectiva mais comercial e outras com um enfoque mais acadêmico. Neste trabalho, utiliza-se uma perspectiva mais acadêmica, em uma classificação realizada por Russell e Norvig (2016), que no livro *Artificial Intelligence: A Modern Approach* apontam os principais subcampos da Inteligência Artificial. A figura 18 apresenta as principais áreas com os seus desdobramentos.

**Figura 18** - Subcampos da Inteligência Artificial




Fonte: Baseado em Russell e Norvig (2016)

A figura 18 destaca seis campos principais da Inteligência Artificial e os seus desdobramentos. O primeiro campo é o de resolução de problemas, que envolve algoritmos de buscas e o *game playing*. Essa é uma das mais tradicionais áreas da Inteligência Artificial, tratando de como resolver problemas que tradicionalmente humanos resolvem, aplicando habilidades cognitivas, como reconhecimento de padrões e buscas (CHIJINDU, 2012).

Outra tradicional área se refere ao conhecimento e ao raciocínio. Esse campo reúne agentes que regem logicamente, que partem de conhecimento do mundo para realizar a tomada de decisão, além de envolver a questão de inferir a partir do conhecimento existente. Além disso, existem sistemas computacionais que são baseados no raciocínio lógico (RUSSELL; NORVIG, 2016).

O terceiro subcampo é o de ação lógica, que trata dos agentes de planejamento. Esses agentes são similares aos algoritmos de resolução de problemas, mas tem um aspecto mais direto, utilizando representações e lógicas mais explícitas.

O campo de conhecimento e raciocínio incerto trata da questão da incerteza e de sistemas probabilísticos de raciocínio. Essa é uma importante área da Inteligência Artificial, pois trata de um aspecto natural das pessoas e de sua comunicação. A ideia de se trabalhar com a incerteza, é, portanto, essencial para que a interpretação e a compreensão dos aspectos não computacionais sejam efetivas. (LI; DU, 2017).

O quinto subcampo trata da comunicação, percepção e ação, sendo um dos mais relevantes para este trabalho. Nesse subcampo encontram-se as pesquisas de Inteligência Artificial vinculadas à linguagem natural, com o seu processamento e compreensão. Ademais, destacam-se trabalhos que estão buscando criar robôs que simulam a linguagem humana, que tem se popularizado nos últimos anos.

Por fim, o último campo trata do aprendizado. Esse campo é um dos mais conhecidos e fundamentais para o desenvolvimento desta tese. Nesse campo estão presentes as pesquisas que tratam de aprendizado de máquina (*machine learning*) e redes neurais, com as suas diferentes formas. O aprendizado busca criar sistemas de alta performance que tenham um alto nível de autonomia, diferentemente dos sistemas que não utilizam Inteligência Artificial. Em suma, o campo do *machine learning* busca criar programas que aprendem a partir das experiências que esse sistema vivencia.

No âmbito deste trabalho, dois subcampos da Inteligência Artificial colocam-se em destaque: processamento de linguagem natural e *Machine Learning*. Ambos serão explorados, respectivamente, na sequência.

## 4.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Dentro da área de Inteligência Artificial, encontra-se a área de processamento de linguagem natural, conceito fundamental para o desenvolvimento teórico deste trabalho. O processamento de linguagem natural será utilizado posteriormente, para a realização do modelo e para a integração que será proposta entre Web Semântica e processamento de linguagem natural, além de ser importante para explicar o modo de recuperação da informação que será descrito nesta tese.

### 4.3.1 CONCEITO DE PROCESSAMENTO DE LINGUAGEM NATURAL

A área de processamento de linguagem natural (PLN) se encontra posicionada dentro da área de Inteligência Artificial, por fazer uso dos mecanismos desse segundo campo de estudo, para realizar a interpretação e a compreensão dos textos.

Vieira e Lopes (2010, p. 184) apontam que: “Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais.”

Essa definição insere a área de PLN dentro da Ciência da Computação, tendo, porém, uma série de interdisciplinares, em especial com a linguística, pela necessidade de compreensão e tratamento da linguagem natural e humana.

Uma outra definição dada por Liddy (2001, p. 1, tradução nossa), afirma que

> O Processamento de Linguagem Natural é uma gama teoricamente motivada de técnicas computacionais para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística com a finalidade de obter processamento de linguagem semelhante ao humano para uma série de tarefas ou aplicações.

A definição clássica, dada pela autora, destaca a união entre as técnicas computacionais e a linguística, visando a obter um tratamento da linguagem semelhante ao obtido pelos humanos. Como relatado anteriormente, o uso da Inteligência Artificial nesse contexto é necessário para que, assim como os humanos, o processamento consiga identificar significado e contexto das palavras.

A autora explica alguns dos elementos inseridos nessa definição, apontando que:

- Textos que ocorrem naturalmente: textos de quaisquer gêneros ou idiomas, que estão expressos de forma escrita ou oral, que não são construídos com propósito de serem analisados, ou seja, devem ser coletados do mundo real (LIDDY, 2001);

- Um ou mais níveis de análise linguística: há diversos níveis de análise linguística, que provavelmente são todos utilizados pelos humanos ao se comunicarem. No entanto, os sistemas de PLN podem utilizar níveis de análise linguística distintos, ou combinações entre esses níveis, podendo um sistema ter características de um PLN forte ou de um PLN fraco. (LIDDY, 2001).
- Processamento de linguagem semelhante ao humano: isso revela que o PLN se encontra dentro da área da Inteligência Artificial, porém esse campo depende de diversas outras disciplinas, para que se obtenha um desempenho semelhante ao de um humano. (LIDDY, 2001).
- Série de tarefas ou aplicações: indica que o PLN não é uma meta em si, e deve estar vinculado a outras tarefas para se atingir determinados objetivos. Exemplos seriam sistemas de recuperação da informação, sistemas de pergunta e respostas, tradução automática, entre outros. (LIDDY, 2001).

A autora relata ainda que o PLN tem como objetivo:

> O objetivo da PNL [...] é ‘realizar um processamento de linguagem semelhante ao humano’. A escolha da palavra "processamento" é muito deliberada e não deve ser substituída por "compreensão". Pois embora o campo da PNL tenha sido originalmente referido como Natural Language Understanding (NLU) nos primeiros tempos da IA, está bem acertado hoje que embora a meta da PNL seja a verdadeira NLU, essa meta ainda não foi alcançada. Um sistema NLU completo seria capaz de: 1. Paráfrase de um texto de entrada; 2. Traduzir o texto para outro idioma; 3. Responder a perguntas sobre o conteúdo do texto; 4. Tirar inferências do texto. (LIDDY, 2001, p. 2, tradução nossa).

A visão da autora destaca alguns pontos importantes acerca do PLN. O primeiro está em que há uma diferença entre o PLN e o entendimento de linguagem natural (*Natural Language Understanding* - NLU), apesar do objetivo do PLN ser o entendimento da linguagem. Adicionalmente, a autora destaca alguns pontos que o NLU deve ter, que também deve existir em um sistema de PLN, como parafrasear texto, traduzir, responder perguntas e tirar inferências.

Uma outra definição apresenta uma visão sobre como o PLN se desenvolveu, e como pode ser visto hoje. Silva (2007, p. 4) afirma que “[...] o PLN apresenta-se como um campo de estudos bastante heterogêneo e fragmentado, acumulando uma vasta literatura e agregando pesquisadores das mais variadas especialidades, com formação acadêmica, embasamento teórico e interesses também bastante diversos.”

Essa definição evidencia que a área de PLN abrange diversos estudos, e uma literatura muito ampla, que não está restrita a uma única área de estudo. Nesse sentido, vale destacar que

há diversos textos nas áreas de Ciência da Computação, Inteligência Artificial, Psicologia, Lógica, Linguística, Computação Cognitiva, e, mais recentemente, Ciência da Informação e Biblioteconomia tratando da temática do PLN.

Uma outra clássica definição de PLN é dada por Joshi (1991, tradução nossa) que afirma que essa área: "[...] se preocupa com (i) o estudo de modelos matemáticos e computacionais da estrutura e função da linguagem, seu uso e sua aquisição e (ii) o projeto, desenvolvimento e implementação de uma ampla gama de sistemas, como mencionado acima."

O autor insere um importante elemento, o dos modelos matemáticos que são necessários para a realização do processo de transposição da linguagem natural para o computador. Além disso, o autor destaca que as implementações e aplicações realizadas no âmbito dos estudos de PLN fazem parte desse campo de estudos.

Nunes (2008) aponta uma importante questão acerca do PLN, relatando que essa área por vezes é confundida, ou tratada como igual às áreas de Linguística Computacional e Linguística Aplicada, ou mesmo com outras áreas da Inteligência Artificial, como *Text Mining*. Visando a diferenciar a área de PLN das demais, a autora considera que:

> De modo geral, em PLN buscam-se soluções para problemas computacionais, ou seja, tarefas, sistemas, aplicações ou programas, que requerem o tratamento computacional de uma língua natural (português, inglês, etc.), quer seja escrita (texto) ou falada (fala). Línguas naturais alternativas, como a linguagem de sinais para os deficientes auditivos, têm igualmente sido alvo crescente de estudos para alguma forma de automatização. [...] Por esse motivo, PLN é quase sinônimo de processamento de língua escrita. (NUNES, 2008, p. 3).

A visão apresentada pela autora indica a relação entre o tratamento computacional e a linguagem natural, especialmente no que tange a língua escrita. Além de indicar que as pesquisas realizadas com linguagem falada e com a linguagem de sinais estão presentes dentro da área de PLN.

Luz (2013, p. 3) sumariza essas questões afirmando que: "PLN tem o objetivo geral de processar linguagem humana para que seja compreensível pelo computador." Essa definição situa brevemente o que é e o objetivo do PLN, que gira em torno de permitir que o computador seja capaz de compreender a linguagem humana.

Dessa forma, este trabalho utiliza o termo processamento de linguagem natural, pois, para atingir uma aproximação entre as áreas de Web Semântica e a linguagem natural, é necessário que haja efetivamente um processamento dos termos de linguagem natural, em seus diversos níveis, como será explicado melhor no decorrer desta seção.

Diante dos conceitos relatados, apresenta-se a seguir um breve histórico da área de processamento de linguagem natural.

### 4.3.2 HISTÓRICO DO PROCESSAMENTO DE LINGUAGEM NATURAL

A história do processamento de linguagem natural inicia-se na década de 1940, em que surgiram as primeiras iniciativas de tradução automática (*machine translation* - MT), fazendo uso dos conceitos, que hoje estão enquadrados como parte de PLN.

Liddy (2001, p. 4, tradução nossa) aponta que:

> Weaver e Booth [...] [iniciaram] um dos primeiros projetos de MT em 1946 sobre tradução computacional, baseada na perícia em quebrar códigos inimigos durante a Segunda Guerra Mundial; foi geralmente aceito que foi o memorando de Weaver de 1949 que trouxe a idéia de MT para aviso geral e inspirou muitos projetos [...] Eles sugeriram usar idéias da criptografia e da teoria da informação para tradução de idiomas. A pesquisa começou em várias instituições de pesquisa nos Estados Unidos dali a alguns anos.

O trecho evidencia que a ideia de utilizar a linguagem natural em processamento computacional está presente há vários anos, além de apontar um dos grandes objetivos do PLN que está na tradução de idiomas. Além disso, cabe destacar um dos pontos levantados ao apresentar as principais definições de PLN, que está no apoio que essa área deve dar a outras aplicações.

Silva (2010, p. 6), ao comentar sobre esse cenário apresentado, relata que: “Para eles, traduzir não era diferente de decifrar códigos. A criptografia – técnica que hoje sabemos ser absolutamente inadequada ao tratamento computacional das línguas humanas – era a única ferramenta de que dispunham para criar os programas tradutores.”

Nesse contexto, o tratamento dado à polissemia dos termos, assim como às análises das unidades linguísticas já se mostrava problemático, era um dos principais desafios da época e continua sendo investigado até os dias atuais.

Na década de 1950, foram desenvolvidas novas iniciativas que começaram a aprimorar o processo e a definir efetivamente o que seria o PLN. Um importante movimento nesse caso foi feito por Chomsky (1957), ao publicar o livro *Syntatic Structures*, que inseriu o conceito de gramática gerativa, dando uma melhor visão de como a linguística poderia ajudar no processo de tradução automatizada.

Anteriormente a isso, em 1952, em uma iniciativa de várias universidades, tais como o MIT e a Universidade de Georgetown, foi apresentado um sistema capaz de traduzir do russo para o inglês cinquenta frases tratando da temática de química. Para isso, criou-se um dicionário de 250 palavras, além de uma gramática escrita com apenas seis regras do russo. Essa iniciativa teve bastante destaque, pois obteve relativo sucesso, apresentando um grande avanço para as

iniciativas de tratamento de linguagem natural. Apesar desse sucesso, a qualidade da tradução ainda era bem baixa, traduzindo termo por termo, sem passar por nenhum tipo de análise sintática. (SILVA, 2010).

Posteriormente, em 1966, a ALPAC (*Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council*) lançou um relatório que afirmava que as iniciativas realizadas de tradução automatizada não eram eficientes e, assim, boa parte dos trabalhos focados em PLN foram paralisados. Destaca-se que, de fato, a maioria dos trabalhos realizados não tinham nenhum embasamento linguístico, o que deixou em descrédito toda a área de PLN. (LIDDY, 2001).

Após esse relatório, demoraram alguns anos para se ter novamente pesquisas que apresentassem alguma evolução na área de processamento de linguagem natural. Em especial, na tese de doutorado de Winograd, foi proposto o sistema chamado SHRDLU, no ano de 1970, considerado um marco para os estudos de PLN. Esse sistema simulava um robô, que manipulou blocos em uma mesa. (LIDDY, 2001; SILVA, 2010).

O trabalho SHRDLU foi considerado um grande avanço, demonstrando possibilidades de realizar conversações com máquinas, apresentando uma nova forma de interação humano-computador, que ocorria por meio da fala. Tal trabalho conseguiu provar que era possível o entendimento da linguagem natural por meio de um computador.

Na década de 1980, novos trabalhos trouxeram contribuições para a área de PLN, voltando a utilizar abordagens não-simbólicas, como o uso das estatísticas para apoiar o processamento da linguagem. Tais abordagens tinham perdido popularidade no início do PLN, e foram importantes para complementar as abordagens simbólicas, aprimorando os resultados alcançados com as técnicas aplicadas. (LIDDY, 2001).

Posteriormente, na década de 1990, o campo avançou significativamente, apoiado pelos avanços tecnológicos e da área de computação, em especial da capacidade de processamento e de memória, pela grande quantidade de textos disponíveis eletronicamente na rede e pelo advento da internet e da Web. Nessa década, as abordagens estatísticas ajudaram a tratar a maioria dos problemas genéricos na linguística computacional. (LIDDY, 2001).

Nas duas décadas seguintes, 2000 e 2010, o desenvolvimento de técnicas e sistemas utilizando técnicas de PLN aumentaram ainda mais, especialmente pelo grande avanço da Inteligência Artificial. Esse avanço, potencializado pelas técnicas de aprendizado de máquinas, permitiu que o PLN se popularizasse e passasse a ser utilizado em uma série de aplicações, para compreender textos dos mais diversos tipos.

Destacam-se iniciativas como o IBM Watson, que congrega diversas iniciativas de aprendizado de máquinas, inclusive de processamento de linguagem natural. O IBM Watson foi lançado em 2010, e aumentou significativamente a escala na qual essas aplicações passaram a ser utilizadas.

O IBM Watson apresenta atualmente quatro bibliotecas principais para o tratamento de PLN: *Watson Natural Language Understanding*, exposta como: "Analise o texto para extrair metadados do conteúdo, como conceitos, entidades, palavras-chave, categorias, sentimento, emoção, relações e papéis semânticos." (IBM, 2019, tradução nossa); *Watson Discovery*, apresentada como: "Desbloqueie o valor oculto nos dados para encontrar respostas, monitorar tendências e padrões de superfície com o mecanismo de visualização nativa mais avançado do mundo." (IBM, 2019, tradução nossa); o *Watson Knowledge Studio*, mostrado como: "Ensine ao Watson a linguagem do seu domínio com modelos personalizados que identificam entidades e relacionamentos exclusivos do seu setor em textos não estruturados" (IBM, 2019, tradução nossa); e *Watson Natural Language Classifier*, apresentado como: "Classificação de texto facilitada. Use o aprendizado de máquina para analisar texto e rotular e organizar dados em categorias personalizadas." (IBM, 2019, tradução nossa).

O Watson influenciou diversas iniciativas com o mesmo foco, como bibliotecas de processamento de linguagem natural da Amazon Web Services e Microsoft Azure. Essas iniciativas são de fácil uso e favorecem a integração com diversos serviços, podendo ser utilizadas para sistemas realizarem, por exemplo, análise de sentimentos, classificação de termos, entre outros.

A seguir, a figura 19 apresenta como ocorreu o avanço da área de PLN, desde a década de 1950, até os dias atuais, da década de 2010. Silva (2010) apresenta esse avanço partindo da década de 1950, chegando até os anos 1990. Além destas décadas, inseriu-se as décadas de 2000 e 2010.

**Figura 19** - Avanço do processamento de linguagem natural

**Década de 1950:**
A tradução automática  sistematização computacional das classes de palavras da gramática tradicional  identificação computacional de poucos tipos de constituintes oracionais
**Década de 1960:**
Novas aplicações e criação de formalismos  primeiros tratamentos computacionais das gramáticas livres de contexto  criação dos primeiros analisadores sintáticos  primeiras formalizações do significado em termos de redes semânticas
**Década de 1970:**
Consolidação dos estudos do PLN  implementação de parcelas das primeiras gramáticas e

analisadores sintáticos  busca de formalização de fatores pragmáticos e discursivos
**Década de 1980:**
Sofisticação dos sistemas  desenvolvimento de teorias linguísticas motivadas pelos estudos do PLN
**Década de 1990:**
Sistemas baseados em "representações do conhecimento"  desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos linguísticos e extralinguísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos linguísticos
**Década de 2000:**
Avanço da Inteligência Artificial e algoritmos de aprendizado de máquinas, que apoiaram o desenvolvimento de PLN de forma mais ativa e popularizada.
**Década de 2010:**
Serviços como o IBM Watson revolucionaram o modo como PLN foi utilizado nos mais diversos serviços, passando a ser empregado em diversos serviços e soluções computacionais, permitindo a inserção de PLN em diversos sistemas e aplicações.

Fonte: adaptado de Silva (2010, p. 8).

A figura 19 revela uma síntese da evolução do PLN, em que é possível ver a transformação, saindo das iniciativas de tradução automatizada, até o aprimoramento e a popularização do PLN com os algoritmos de aprendizado de máquina.

Uma das principais formas de tratar a área de PLN está vinculada aos níveis que são utilizados nesse campo de estudo. Apresentam-se na sequência tais níveis.

### 4.3.3 NÍVEIS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Há diversas abordagens para apresentar o processamento de linguagem natural, que pode ser classificado por suas formas e modelos. Uma abordagem bastante tradicional diz respeito aos níveis de PLN, que demonstram qual a capacidade de processamento, levando em consideração os aspectos que os humanos utilizam para compreender um texto. Destaca-se que o processo de compreensão é bastante dinâmico, havendo uma grande interação entre esses níveis de processamento. Assim, quanto maior for a capacidade e o nível alcançado por um sistema de processamento de linguagem natural, melhor serão os resultados obtidos.

Nesse sentido, destacam-se os seguintes níveis de PLN: fonologia, morfologia, léxico, sintático, semântico, discursivo e pragmático. (LIDDY, 2001), que serão apresentados a seguir.

#### 4.3.3.1 Fonologia

A fonologia trata “[...] do relacionamento das palavras com os sons que produzem.” (GONZALEZ; LIMA, 2003, p. 3). Em uma das definições do dicionário, o termo fonologia é apresentado como: “Parte da linguística que estuda os sons a partir do ponto de vista da língua, do sistema e da forma, compreendendo o som como uma unidade de língua virtual, relevante, significativa e portadora de valor determinado. (A fonologia avaliza os padrões sonoros de uma língua específica e, para fazê-lo, determina quais sons fonéticos são significativos e explica o modo como esses sons são interpretados pelo falante nativo.)” (MICHAELIS, 2019).

Em uma perspectiva da linguística, a Fonologia é: “[...] a área da linguística que se ocupa da descrição e análise da massa amorfa fônica ou gestual. E a fonologia é a área de linguística que se ocupa da descrição e análise dos significantes de cada língua, ou seja, da porção que cada língua formatou a partir da massa amorfa fônica ou gestual.” (VIOTTI, 2008, p. 45).

Em síntese, a fonologia irá atuar no tratamento e na interpretação dos sons das palavras, para compreender os elementos fonéticos relevantes. A fonologia/fonética é o primeiro passo para a interpretação de linguagem natural, quando obtida por meio da voz, pois, é necessário que haja a compreensão dos sons, transpondo-os em palavras.

No contexto do PLN, há três regras principais para a análise fonética:

1. Regras fonéticas: focadas em palavras com sons (LIDDY, 2001);
2. Regras fonéticas: focadas em variações de pronúncias, quando as palavras são faladas juntas (LIDDY, 2001);
3. Regras prosódicas: focadas nas flutuações de entonação e ênfase dado por meio de uma sentença (LIDDY, 2001).

Destaca-se que estas regras estão todas focadas em interpretar os sons das suas palavras, tratando dificuldades como pronúncia, entonação e ênfase, que torna o processo complexo, mas necessário, para realizar o tratamento da fonologia das palavras.

#### 4.3.3.2 Morfologia

O próximo aspecto é a morfologia. A análise morfológica aprofunda a compreensão dos termos, abordando aspectos de sua constituição. Na linguística, a morfologia é apontada como a área da linguística que estuda a palavra. (VIOTTI, 2008).

Outra definição é dada no dicionário, com a morfologia sendo apontada como: “5. Estudo das diversas classes de palavras, seus paradigmas de flexões e suas exceções. [...] 7.

Parte da linguística que trata das estruturas e dos processos de formação das palavras." (MICHAELIS, 2019).

A definição apresentada está focada na gramática e na linguística, em que é possível verificar que a morfologia está orientada para o modo como as palavras são formadas, e aspectos relacionados a isso. Nesse sentido, Gonzalez e Lima (2003, p. 3) afirmam que a morfologia trata "[...] da construção das palavras a partir de unidades de significado primitivas e de como classificá-las em categorias morfológicas."

Os autores demonstram, por meio dessa definição, que se trata de um aprofundamento do léxico, visto que, com as análises morfológicas inicia-se o processo de tratamento e classificação em categorias, além de inserir a questão de unidades de significados, ainda que de forma incipiente. Essas unidades são chamadas de morfemas. O morfema é o menor elemento de significado, e são elementos desse tipo que compõem as palavras. Existem seis tipos de morfemas: desinência, raiz, radical, afixo, tema e vogal temática.

Um exemplo disso é a palavra desconsideração. Essa palavra é composta pelo prefixo *des*, pela raiz *considera* e pelo sufixo *ção*. Nesse sentido, as pessoas são capazes de compreender o significado das palavras, pelo entendimento dos morfemas que fazem parte de uma palavra.

Similarmente ao processo realizado por uma pessoa, as técnicas de processamento de linguagem natural buscam compreender o significado de um termo utilizando os morfemas. Liddy (2001, p. 7, tradução nossa) afirma que: "[...] um sistema de PNL pode reconhecer o significado transmitido por cada morfema, a fim de ganhar e representar o significado." Um exemplo disso na língua inglesa está em toda vez que for adicionado o sufixo *ed* em um verbo, isso implica que a ação desse verbo aconteceu no passado. Essa informação é essencial para compreender o significado de um texto.

#### 4.3.3.3 Léxico

Aprofundando o nível de compreensão, o próximo nível é o léxico. O nível léxico trata de alguns significados que os termos podem ter, sendo necessário para os sistemas tratarem o significado das palavras de forma individual.

O termo léxico é definido como:

> 3 Conjunto total das palavras com características sintáticas de que dispõe determinado idioma; composição lexical. 4 Lista de palavras, com suas respectivas definições, usadas de maneira peculiar ou com sentido diferente do comum, por um autor ou por um grupo de pessoas, ou usada num período, num movimento etc. (MICHAELIS, 2019)

A definição do dicionário deixa clara a noção de haver extensa lista de palavras com suas respectivas definições, cuja totalidade constitui o léxico de determinada língua.

No âmbito de sistemas, podem ser realizados diversos tipos de processamentos. Uma forma estaria em inserir *tags* que representam uma parte da fala para cada palavra tratada, visando contribuir para a compreensão do termo, de acordo com o contexto que ele foi inserido. Outra forma, estaria na substituição do termo por um outro que tenha uma maior representação semântica, no caso daqueles termos que possuem um único sentido ou significado. (LIDDY, 2001).

Uma representação léxica pode ser vista na figura 20, em que a palavra "lancha" (*launch* em inglês) é apresentada por meio de predicados lógicos.

**Figura 20** - Predicados lógicos do termo "lancha"

| |
|---|
| lancha (um barco grande usado para transportar pessoas em rios, etc.)<br><br>((CLASSE BARCO) (PROPRIEDADES (GRANDE) (OBJETIVO (PREDICAÇÃO (CLASSE CARGUEIRO) (OBJETO PESSOAS))))) |

Fonte: adaptado de Liddy (2001, p. 8).

Por meio da figura 20, é possível visualizar que o termo foi decomposto em vários predicados e propriedades. Assim, por meio desse processo, é possível fazer uma série de interpretações, com um certo nível de complexidade, semelhantemente ao que pessoas conseguem realizar. Nesse exemplo, tem-se propriedades sobre o tipo de elemento, sinônimos e significados dos termos que, nesse caso, apontam que esse termo é da classe BARCO, tendo a propriedade GRANDE, tendo o objetivo de ser um CARGUEIRO de PESSOAS. Todo esse processo ocorre por meio de predicados lógicos, tornando o processo mais facilmente compreensível.

#### 4.3.3.4 Sintático

O próximo nível aprofunda o modo como se tratam os termos em sistemas de processamento de linguagem natural. Nesse nível, os termos são tratados de forma relacionada, em que há combinação entre os termos para determinar uma frase.

Nesse contexto, Viotti (2008, p. 57-59) aponta que:

> A sintaxe é a área da gramática que trata da estrutura da sentença. [...] A sintaxe se ocupa, justamente, de estudar as propriedades de combinação de certas expressões linguísticas. São essas propriedades que determinam, em grande parte, a construção e a estruturação das sentenças de uma determinada língua.

O termo sintaxe é definido como:

> 1 Parte da gramática que trata da disposição das palavras na frase, da relação entre essas palavras, bem como das combinações e das relações lógicas das frases no enunciado. 2 Cada um dos elementos da estrutura linguística que determina as relações entre os componentes da oração. (MICHAELIS, 2019).

As definições demonstram que, ao se tratar de sintaxe, o foco está no relacionamento e na combinação entre os termos, que permite a relação entre os componentes de uma oração. Uma outra definição é dada por Gonzalez e Lima (2003, p. 3), que afirma que o sintático está no: "[...] relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças."

Por meio da definição, verifica-se que, comparado aos outros níveis, o nível sintático permite que frases e sentenças possam ser constituídas, tendo uma sequência lógica. No nível humano, as pessoas criam frases seguindo regras sintáticas, em que as palavras assumem determinados papéis, criando assim uma oração, que tenha sentido gramatical.

Em nível de sistema, requer-se uma gramática e um analisador para realizar o processamento. Em suma, a saída de um analisador sintático é uma representação da sentença, contendo as relações de dependência entre os termos. Destaca-se que existem diversas gramáticas que podem ser utilizadas, impactando no tipo de analisador que deverá ser usado. (LIDDY, 2001).

Liddy (2001, p. 8, tradução nossa) destaca ainda que:

> Nem todos os aplicativos de PNL exigem uma análise completa das sentenças; portanto, os desafios restantes na análise do escopo de definição de anexo e conjunção de frase preposicional não impedem mais as aplicações para as quais as dependências frasais e de cláusula são suficientes. A sintaxe transmite significado na maioria das linguagens porque a ordem e a dependência contribuem para o significado.

A questão relatada pela autora demonstra a importância da análise sintática, uma vez que a sintaxe expressa um nível de significado, em que a ordem dos termos e a dependência entre eles tem impacto significativo na compreensão desses termos. Adicionalmente, a autora aponta que o PLN deve se concentrar em algumas dessas questões ao realizar a análise.

Para compreender o que a autora apontou, observam-se duas frases: “A mulher quebrou o martelo” e “O martelo quebrou a mulher”, que apresentam as mesmas palavras, diferindo apenas em termo de sintaxe, o que resulta em significados completamente diferentes.

#### 4.3.3.5 Semântica

O próximo nível de análise é o semântico, que aprofunda as relações entre as palavras, além de apontar os seus significados. A linguística aponta que: “O estudo do significado é feito pela semântica e pela pragmática. [...] De maneira geral, a Semântica trata da significação linguística independentemente do uso que se faz da língua.” (VIOTTI, 2008, p. 62-65).

O trecho destacado demonstra que a questão do significado é tratada tanto pela semântica, quanto pela pragmática. A diferença está que a semântica não leva em consideração o contexto em que a língua foi utilizada, diferentemente da pragmática, que será tratada com detalhes posteriormente.

Vale destacar que, o termo ‘semântica’ é definido como: “1 Ramo da linguística que estuda a significação das palavras e suas mudanças de sentido ao longo do tempo, bem como a representação do sentido dos enunciados. 2 O significado dos vocábulos, por oposição à sua forma.” (MICHAELIS, 2019).

Em ambas as definições, verifica-se que o foco da semântica está no significado das informações, além de apontar a questão do sentido de um enunciado. Quando pensado sob a perspectiva humana, a semântica está na interpretação do sentido dos termos, de forma individual e relacionada a outros termos.

Esse ponto demonstra a complexidade do tratamento da semântica por meio de técnicas de processamento de linguagem natural, devido a subjetividade que contempla o aspecto do significado dos termos. Nesse sentido, Gonzalez e Lima (2003, p.3) apontam que o nível semântico trata “do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças.”

Vale destacar o seguinte aspecto: não é somente pela semântica que se tem o significado dos termos, pois é necessário que todos os outros níveis sejam implementados para que esse aspecto possa existir. No nível de sistema, o processamento semântico irá determinar os possíveis significados de uma determinada frase ou conjunto de frases, focando principalmente na interação entre os termos dentro da sentença.

Liddy (2001) aponta um importante aspecto do processamento semântico, que está ligado a desambiguação semântica, quando se tem palavras com diversos significados. Esse processo é importante para que seja escolhido um único significado para uma palavra, e consequentemente, para uma frase. A autora destaca que se podem aplicar diversas técnicas para a realização da desambiguação semântica.

Para compreender o que autora afirmou, tome-se, por exemplo, uma frase com a palavra "vela". Vela pode significar a vela de um barco, a vela feita de cera para iluminar um ambiente, ou ainda, a 3ª pessoa do singular do presente do indicativo (ele ou ela vela) e segunda pessoa do singular do imperativo afirmativo (vela tu) do verbo velar, que significa estar atento. Nesse caso, se as demais palavras da sentença forem necessárias para realizar o processo de definição do sentido, será preciso realizar uma desambiguação semântica.

#### 4.3.3.6 Discursivo

O nível discursivo é o penúltimo nível, e um dos mais complexos, pois, envolve uma série de elementos, em se tratando de uma manifestação concreta. Nesse sentido, o termo discurso, no que se refere à linguística, pode ser definido como: "Manifestação concreta da língua. As 10 classes gramaticais em que se enquadram as palavras. 5. Comunicação oral ou escrita que pressupõe um locutor e um interlocutor. 6 Reprodução de palavras atribuídas a outra pessoa." (MICHAELIS, 2019).

As definições apresentadas revelam a complexidade que o nível discursivo possui, uma vez que está vinculado à questão da manifestação concreta da língua, envolvendo os problemas de comunicação, oral e escrita, pressupondo as figuras do locutor e interlocutor, estando em um contexto mais amplo que o de sentenças isoladas.

No âmbito do processamento de linguagem natural, o nível discursivo irá tratar de contextos maiores, contemplando mais do que uma sentença, enquanto a sintaxe e a semântica irão tratar de uma única sentença. Liddy (2001, p. 9, tradução nossa) aponta que:

> [...] ele não interpreta textos de multi sentenças apenas com sentenças concatenadas, cada uma das quais pode ser interpretada individualmente. Em vez disso, o discurso concentra-se nas propriedades do texto como um todo, que transmite significado fazendo conexões entre sentenças componentes.

A autora destaca que o foco do nível discursivo está na compreensão total de um texto, interpretando outros tipos de elementos. Destaca-se que esse nível necessita tanto da análise sintática quanto semântica, para permitir tal análise discursiva. Ressalta-se a complexidade desse tipo de análise, devido à necessidade de interpretar as relações entre as diversas sentenças.

Os dois métodos principais para realizar a análise discursiva são 1) resolução de anáforas; e 2) reconhecimento de estrutura de texto e discurso. O primeiro método trata da substituição de palavras por pronomes vagos, enquanto o segundo método determina as funções das sentenças de um texto. Um exemplo seria um texto jornalístico, que é dividido em: título, introdução, eventos anteriores, citações, entre outros. (LIDDY, 2001).

#### 4.3.3.7 Pragmática

A pragmática é o último nível do processamento de linguagem natural, sendo também o mais complexo, pois está vinculado às questões do contexto em que os termos se encontram.

Pragmática, no contexto da linguística, pode ser definida como: "A Pragmática [...] teria como objeto o estudo da significação construída a partir do momento em que a língua é posta em uso, ou seja, em uma determinada situação de fala." (VIOTTI, 2008, p. 65).

Outra definição considera-a como:

> Ramo da linguística que trata do uso da linguagem do ponto de vista do falante, levando em conta suas motivações psicológicas, o efeito que sua fala causa no interlocutor, o modo como certos fatores podem interferir na escolha das formas linguísticas e no nível de formalidade etc. (MICHAELIS, 2019)

As definições demonstram a complexidade do pragmatismo na linguística, pois envolve uma série de questões subjetivas e que são difíceis de serem compiladas e tratadas, pelo fato de estarem vinculadas ao contexto em que a frase foi dita ou escrita. De forma mais simples, Gonzalez e Lima (2003, p. 3) apontam o nível pragmático como o que trata "[...] do uso de frases e sentenças em diferentes contextos, afetando o significado." Os autores trazem esse importante elemento da pragmática, que é o estar vinculada ao contexto, o que afeta o significado, para além da semântica. Assim, tem-se outro elemento a ser tratado durante os processos de PLN.

Nesse âmbito, a análise pragmática deve considerar elementos referentes ao uso intencional da linguagem, utilizando o contexto, além do significado dos termos, para compreender o sentido das sentenças. Liddy (2001, p. 9, tradução nossa) afirma que: "O objetivo é explicar como o significado extra é lido em textos sem estar realmente inserido neles. Isso requer muito conhecimento geral, incluindo a compreensão de intenções, planos e objetivos."

O aspecto trazido pela autora relata que os elementos de PLN devem considerar conhecimento geral, e ter uma compreensão dos planos e objetivos que um determinado texto apresenta. Dessa forma, o processo deve envolver bases de dados que possam fornecer esse tipo de informação. A autora complementa relatando que o processamento de PLN pode utilizar bases de conhecimento e módulos de inferência nesse processo (LIDDY, 2001).

A figura 21 apresenta um exemplo em que é necessário o nível de pragmática para compreender o texto.

**Figura 21** - Exemplos de análise pragmática

| |
|---|
| Os vereadores da cidade recusaram aos manifestantes uma permissão porque **eles** temiam a violência.<br><br>Os vereadores da cidade recusaram aos manifestantes uma permissão porque **eles** defendiam a revolução. |

Fonte: Adaptado de Liddy (2001, p. 9).

A compreensão das duas frases da figura 21 depende de como o termo *eles* é tratado. De modo que, para compreender o significado do termo *eles,* é necessário que se tenha um conhecimento do contexto, e de questões externas e gerais. Assim, o único modo de compreender o sentido dessa frase, é por meio de uma análise pragmática que considera elementos de conhecimentos gerais.

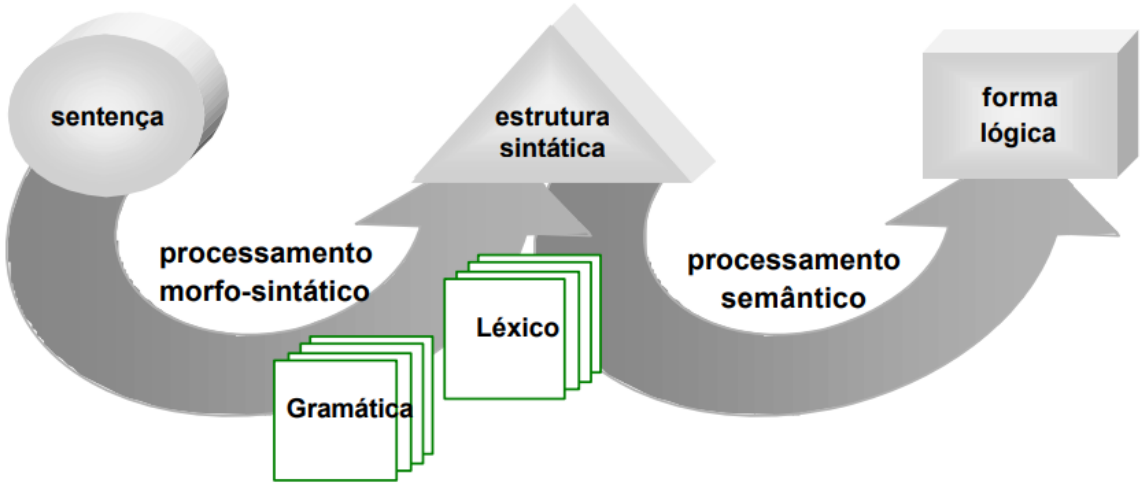
#### 4.3.3.8 Resumo dos níveis

A existência dos diversos níveis não implica que um sistema de processamento de linguagem natural necessite implementar todos eles. Na verdade, a maioria dos sistemas de PLN opta por implementar os níveis mais baixos, pela própria facilidade, e por serem mais assertivos. Além disso, nem sempre é necessário que todos os níveis sejam implementados para que se atinja o objetivo de uma determinada aplicação.

Nesse sentido, um dos pontos que levam ao desenvolvimento de aplicações que utilizam basicamente os níveis mais baixos da linguagem natural está na maior facilidade de tratar as unidades menores de análise, como morfemas e palavras, enquanto os níveis mais altos, devem tratar sentenças e o texto como um todo, o que envolve relações mais complexas.

Um resumo de como esses processos podem ser aplicados em um sistema de PLN pode ser visualizado na figura 22.

**Figura 22** - Tratamento de uma sentença




Fonte: Gonzalez e Lima (2003, p. 3).

Com a figura apresentada, verifica-se que, a partir de uma sentença, realiza-se o processamento morfológico e sintático, utilizando elementos da gramática e léxicos. A partir disso, tem-se uma estrutura sintática, com os elementos e possíveis significados. Em seguida, aplica-se o processamento semântico, obtendo-se assim, os sentidos dos termos, obtendo uma forma lógica, que revela informações mais claras e com significado da sentença inicial, de forma que o computador consiga ter uma melhor visão daquela frase, que inicialmente era apenas linguagem natural.

### 4.3.4 ÁREAS DE APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL

Como relatado por alguns autores, ao definirem processamento de linguagem natural, esse campo está focado em auxiliar outras áreas e campos a atingirem os seus objetivos. Assim, com a popularização e expansão do PLN, há muitos campos e aplicações que estão utilizando as técnicas desse primeiro, aprimorando os processos e aumentando a abrangência de funções.

Além do uso clássico para a recuperação da informação, há diversos outros aspectos como automação de tarefas cotidianas, além dos populares *chatbots*, que estão sendo utilizados em diversos contextos, como pizzarias e assistentes virtuais de bancos e seguradoras. Esse cenário demonstra como o PLN deixou de estar presente apenas em pesquisas acadêmicas, passando a ser utilizado com frequência no cotidiano das pessoas.

Nesse sentido, relatam-se abaixo os principais tipos de aplicações que estão utilizando o PLN.

O primeiro tipo de aplicação diz respeito a sistemas de recuperação da informação que manipulam diretamente bases de dados. Esse cenário vem ganhando força, principalmente, pelas mudanças que estão ocorrendo no modo como as pessoas realizam buscas. Anteriormente, era bastante comum as buscas seguirem uma série de regras, com uma sintaxe bastante específica, ou mesmo acontecerem diretamente utilizando linguagem SQL. No entanto, as novas aplicações de recuperação da informação utilizam uma série de conceitos de processamento de linguagem natural para auxiliar no processo de busca. (NUNES, 2008).

Buscadores como o Google não exigem que o usuário faça a sua busca com o uso de uma sintaxe específica, podendo o usuário inserir palavras e frases que façam sentido para ele. Nesse contexto, o buscador necessita realizar um tratamento de PLN, visando encontrar aquilo que de fato é relevante. Assim, essas aplicações de recuperação da informação estão utilizando os princípios de PLN para realizar as buscas.

Nesse sentido, Nunes (2008) afirma que: “[...] quando a consulta é vista e processada como um fenômeno linguístico, requerendo qualquer tipo de processamento básico, então dizemos que se trata de uma RI linguisticamente motivada e, portanto, de uma aplicação de PLN.” A afirmação da autora traz o cenário de que, quando a busca traz os elementos linguísticos ela torna o próprio processo de busca como uma aplicação de PLN.

A autora considera ainda que há sistemas de recuperação da informação utilizando *cross-language*, em que tanto as consultas quanto os documentos estão em línguas diferentes, e com mais de uma linguagem. Esse é um processo complexo, que envolve técnicas de processamento de linguagem natural, para realizar a tradução e a comparação entre as diversas linguagens.

Ainda vinculado à recuperação da informação, uma outra forma de utilizar o processamento de linguagem natural para apoiar tal processo está no uso de sistemas de pergunta e respostas, vinculados ao *Question Answering*. A ideia está em o usuário escrever uma pergunta, e buscar em um conjunto de bases, a resposta para tal. Há grande complexidade nesse processo, visto que é necessário tratar a pergunta, e encontrar nela, o que o usuário está de fato buscando e, posteriormente, utilizar PLN nos textos que podem conter a resposta, para tratar o que será apresentado como resposta ao usuário. (SILVA et al., 2007).

Vale destacar que o *Question Answering* será um dos principais elementos deste trabalho, e será melhor abordado na seção de recuperação da informação. Nessa seção, será

apresentada a maneira como essa técnica utiliza os elementos de RI, para demonstrar a viabilidade da proposta feita no modelo.

Um outro tipo de aplicação de PLN está no desenvolvimento de sistemas tutores. Em destaque, os chamados sistemas tutores inteligentes fazem uso de diversos conceitos de processamento de linguagem natural. O princípio desse sistema está em uma "rede de conhecimento" com diversos fatos, relações e regras, que permitem ao sistema realizar um diálogo com um indivíduo. Além disso, esse tipo de sistema tem um viés educacional, visando promover instrução imediata aos alunos. (REIS; JAQUES; ISOTANI, 2018).

Complementarmente, Latham, Crockett e McLean (2014) afirmam que os sistemas tutores inteligentes: "[...] constroem um modelo dos objetivos, preferências e conhecimento do aluno, e usam isso para adaptar o ensino ao indivíduo e fornecer assistência inteligente."

Os sistemas tutores inteligentes têm sido cada vez mais utilizados, pelo próprio avanço da área de PLN e de Inteligência Artificial, que tem favorecido o desenvolvimento de tutores que, de fato, consigam estabelecer uma comunicação com uma pessoa. Nesse sentido, vale destacar que os sistemas tutores inteligentes utilizam técnicas de PLN, Inteligência Artificial, aprendizado de máquinas, computação cognitiva, redes bayesianas, entre outros, para simular o comportamento humano. (REIS; JAQUES; ISOTANI, 2018).

Um próximo tipo de aplicação são os sistemas de automação de tarefas. Esse tipo de sistema visa a auxiliar tarefas administrativas e gerenciais de uma empresa, passando por questões como agendamento de reuniões, compras de passagens aéreas e até detecção de erros ortográficos. Silva et al. (2007) apontam a existência dos sistemas SCHED (focado no gerenciamento de agendas de reuniões), GUS (fornecendo informações sobre planejamento de viagens) e o CRITIQUE (tratando de informações de erros ortográficos e gramaticais em documentos administrativos, visando fornecer uma leitura mais fluida).

Vinculado aos dois exemplos anteriores, há um outro tipo de aplicação que são os assistentes virtuais. Esses assistentes virtuais têm o objetivo de auxiliar as pessoas em tarefas cotidianas, que vão desde realizar uma busca, mandar uma mensagem, verificar preços, ou até mesmo realizar uma compra. Destaca-se que, com a popularização dos *smartphones*, os assistentes virtuais passaram a estar presentes, e ser utilizados, por grande parte da população.

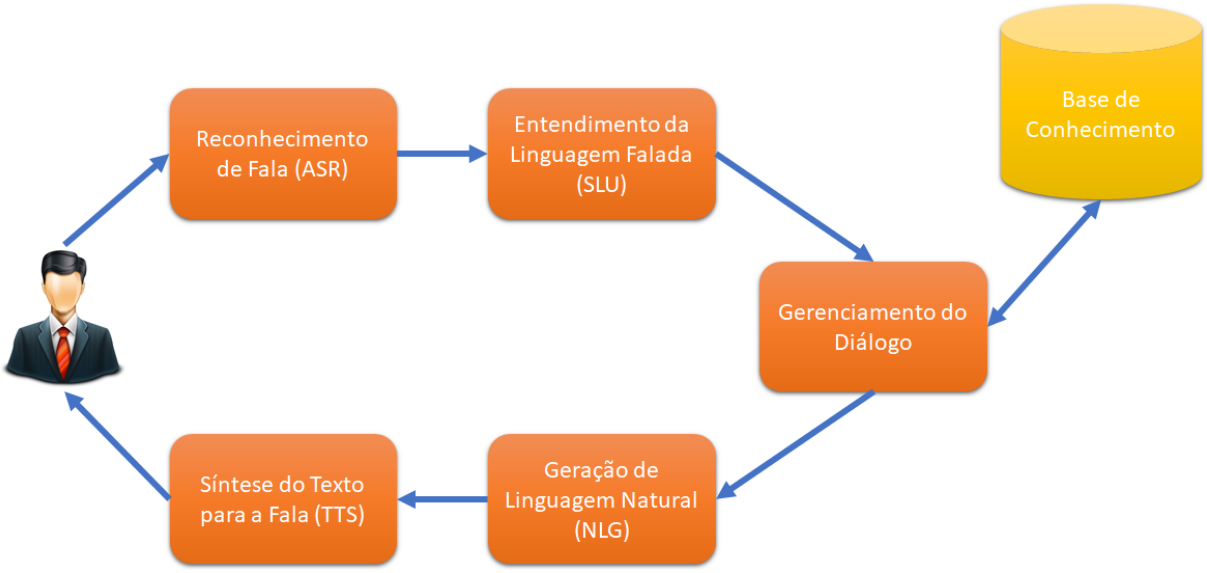
Këpuska e Bohouta (2018, p. 99, tradução nossa) afirmam que:

> Além disso, sistemas de diálogo ou sistemas de conversação podem suportar uma ampla gama de aplicações em empresas, educação, governo, saúde e entretenimento. São os assistentes pessoais, conhecidos por vários nomes, como assistentes pessoais virtuais, assistentes pessoais inteligentes, assistentes pessoais digitais, assistentes móveis ou assistentes de voz.

Os autores citam os exemplos de Microsoft Cortana, Apple Siri, Amazon Alexa e Google Home, como a próxima geração de assistentes virtuais que têm capacidade de auxiliar em diversas áreas, não apenas como assistentes pessoais virtuais. Esses sistemas utilizam diversas técnicas de processamento de linguagem natural, Inteligência Artificial e aprendizado de máquinas, tornando o processo bastante intuitivo, e com uma ampla gama de opções.

Nesse sentido, o processo realizado por esses assistentes ocorre, principalmente, orientado à voz, tanto a forma como a pessoa interage com o sistema, quanto o modo como o sistema responde à pessoa. Isso ocorre, pois, como relata Këpuska e Bohouta (2018), o uso de voz para interação com o sistema será cada vez mais utilizado, em ações como interação com máquinas, televisões, aparelhos de videogame, entre outros. Nesse sentido, a estrutura geral de diálogo, que é a base desses sistemas assistentes virtuais, é apresentada a seguir, na figura 23.

**Figura 23** - Arquitetura base de assistentes virtuais




Fonte: Adaptado de Këpuska e Bohouta (2018, p. 100).

Destaca-se que a arquitetura apresentada tem como foco a fala, realizando o processo de tradução e gerenciamento desse diálogo, visando a atender as demandas dos usuários. Além disso, tem-se a ligação com uma base de conhecimento, que visa a atender as demandas solicitadas pelos usuários.

Um outro tipo de aplicação de processamento de linguagem natural bastante utilizado, são os sistemas de tradução automática. Destaca-se que a tradução automática foi o primeiro tipo de sistema a utilizar PLN. Em síntese, tem-se três tipos de tradução automática: sistemas diretos, sistemas de transferências e sistemas interlinguais. Os sistemas diretos vão buscar a

correspondência direta entre a língua original e a traduzida. Os sistemas de transferências são mais sofisticados, efetuando análise sintática da frase da língua original e, por meio de regras de transferência sintática, constroem a nova frase na língua traduzida. Por fim, os sistemas interlinguais são os mais sofisticados, em que a língua original e a traduzida são intermediadas por uma terceira língua, uma interlíngua, que é uma representação abstrata entre as línguas. (SILVA et al. 2007).

Após demonstrar alguns dos principais tipos de sistemas de processamento de linguagem natural, destacam-se a seguir algumas aplicações que estão embasados nesses tipos de sistemas. Como relatado anteriormente no histórico do PLN, há diversas iniciativas que estão utilizando, atualmente, o processamento de linguagem natural para auxiliar uma grande gama de aplicações, que vão desde tarefas cotidianas, até contextos mais complexos. Vale destacar que áreas como a saúde e a educação estão fazendo uso do PLN para auxiliar as suas tarefas.

### 4.3.5 CLASSIFICAÇÕES DE PROCESSAMENTO DE LINGUAGEM NATURAL

Existem diversas classificações que podem ser aplicadas para o processamento de linguagem natural, que dizem respeito a abordagens, estratégias e modelos que são utilizados no processo de desenvolvimento dos sistemas. Com a evolução dessa área, houve pesquisas que utilizaram diferentes aspectos, ou que combinaram as estratégias para obter resultados satisfatórios. Os exemplos e as aplicações de PLN revelam esse cenário, uma vez que demonstram como a área avançou e se tornou parte do cotidiano das pessoas. Nesse sentido, é necessário compreender os distintos aspectos de PLN, para entender como se aplicam as técnicas atualmente.

A primeira classificação que será apresentada trata das abordagens que podem ser utilizadas no âmbito do PLN.

#### 4.3.5.1 Abordagens de processamento de linguagem natural

Desde o início dos estudos de processamento de linguagem natural, quatro abordagens principais foram utilizadas: simbólica, estatística, conexionista e híbrida.

As duas primeiras abordagens começaram a ser utilizadas desde os primeiros estudos da área de PLN. Já a abordagem conexionista iniciou na década de 1960. Liddy (2001, p. 10, tradução nossa) aponta que:

> Por muito tempo, as abordagens simbólicas dominaram o campo. Nos anos 1980, as abordagens estatísticas recuperaram a popularidade como resultado

> da disponibilidade de recursos computacionais críticos e da necessidade de lidar com contextos amplos e reais. Abordagens conexionistas também se recuperaram das críticas anteriores, demonstrando a utilidade do uso de redes neurais na [área de] PNL.

A afirmação da autora demonstra que as diversas abordagens dividiram prestígio nos diferentes momentos da área de PLN. A seguir apresentam-se as quatro abordagens existentes nesse campo de estudo.

A primeira abordagem é a simbólica. Essa abordagem tem esse nome, pois utiliza especialmente os símbolos linguísticos para realizar as análises e o processamento da linguagem. Assim, é necessária uma profunda análise da linguística, utilizando representações explícitas relativas a esse campo de estudo. Vale destacar que, ao se embasar na abordagem simbólica, o principal instrumento utilizado para realizar as análises são as regras e os elementos léxicos. (LIDDY, 2001).

Nesse contexto, Andreata (2017, p. 20) afirma que: “O método simbólico ou racionalista está baseado no campo da Linguística e faz o uso da manipulação dos símbolos, significados e das regras de um texto.”, mostrando, portanto, que o tratamento e a manipulação dos símbolos são os principais elementos considerados por tal abordagem.

Uma abordagem simbólica tradicional é encontrada nos sistemas baseados em regras, que irá estruturar as informações por meio de proposições lógicas, realizando a manipulação das estruturas por meio de inferências, que visam a determinar as condições verdadeiras para compreender o texto. Liddy (2001, p. 10, tradução nossa) afirma ainda que:

> Os sistemas baseados em regras geralmente consistem em um conjunto de regras, um mecanismo de inferência e um espaço de trabalho ou memória de trabalho. O conhecimento é representado como fatos ou regras na base de regras. O mecanismo de inferência seleciona repetidamente uma regra cuja condição é satisfeita e executa a regra.

Nesse trecho, visualiza-se que o foco dessa abordagem está no mecanismo de inferência, que possibilitará identificar se as regras estão sendo satisfeitas e, assim, possibilitar o processamento dos textos. Ressalta-se que esse processo é possível, pois o conhecimento está representado com fatos e regras, sendo aderente ao mecanismo de inferência.

Uma outra aplicação tradicional da abordagem simbólica está nas redes semânticas. As redes semânticas são estruturas organizadas em nós e ligações, fornecendo uma base de conhecimento para ser utilizada durante o processo de análise dos termos e na realização do processamento de linguagem natural. O uso de redes semânticas é bastante utilizado na área de inteligência artificial, para fornecer informações para as tomadas de decisões e auxiliar os processos de redes neurais.

As redes semânticas refletem, por meio do modo como os conceitos estão interligados, os tipos de relações que existem entre os termos, sendo que, aqueles termos que estão bem próximos, estão ligados por associações mais fortes, enquanto aqueles termos mais distantes possuem ligações mais fracas ou moderadas. No geral, as redes semânticas representam o conhecimento estruturado, sendo também utilizadas no âmbito da Web Semântica, em que há o uso do conceito dessas redes para a estruturação das informações. (LIDDY, 2001).

Por fim, no que tange às abordagens simbólicas, há o método de Brill (1992), que vem sendo utilizado até os dias atuais. Esse método realiza a etiquetagem dos termos como primeira etapa do processo de PLN, definindo se os termos são verbos, substantivos, artigos ou adjetivos. Por exemplo, na frase: "O **almoço** está bom" e a frase "Eu **almoço** todo dia", o termo almoço, considerado isoladamente, poderia ter sido tratado como verbo em ambos os casos, pelo fato de "almoço" ser uma forma verbal pertencente à conjugação do verbo almoçar, ignorando a possibilidade de considerá-lo um substantivo, ou vice-versa: ser considerado como substantivo nos dois casos, ignorando ser uma forma verbal.

Para resolver isso, o método apresentado busca compreender o contexto no qual os termos se encontram, de modo que, ao verificar que o termo **almoço** na primeira frase é precedido do artigo **o**, passa-se a tratar esse termo como um substantivo. Na verdade, esse método se embasa em três regras principais, que buscam tratar esses problemas, considerando o conjunto de termos lexicais, regras gramaticais e, assim, são corrigidos os possíveis problemas realizados na classificação inicial.

Vale destacar que o método apresentado demonstra um pouco do modo como as técnicas da abordagem simbólica se comportam, uma vez que os símbolos são os principais elementos considerados no processo de tratamento de PLN. Destaca-se que os níveis de análise mais utilizados são o léxico e o sintático, sendo assim, uma forma de realizar o tratamento para a linguagem natural.

A outra abordagem amplamente utilizada no processamento de linguagem natural é a abordagem estatística. Tal abordagem começou a ser utilizada já nos princípios do PLN, sendo um dos cernes desse campo de estudo. Andreata (2017) relata que esse método utiliza grande quantidade de textos, buscando, nesses padrões, associações que possam contribuir para as análises de PLN. O autor aponta que o objetivo ao usar essa grande quantidade de informações, é encontrar ou não alguma relação sintática ou semântica entre os termos.

As abordagens estatísticas buscam desenvolver modelos aproximados e generalizados dos elementos e fenômenos linguísticos, tendo como objetivo encontrar exemplos aplicáveis desses fenômenos no corpo do texto. Liddy (2001, p. 11, tradução nossa) complementa

relatando a diferença entre a abordagem estatística e a abordagem simbólica: “Em contraste com as abordagens simbólicas, as abordagens estatísticas usam dados observáveis como a principal fonte de evidência.”

Um tradicional modelo de abordagem estatística é o *Hidden Markov Model* (HMM), que tem sua origem associada aos estudos da fala. Khurana et al. (2017, p. 13, tradução nossa) afirmam que: “Um HMM é um sistema em que um deslocamento ocorre entre vários estados, gerando símbolos de saída viáveis com cada comutador. Os conjuntos de estados viáveis e símbolos únicos podem ser grandes, mas finitos e conhecidos.”

Em suma, o HMM é um autômato finito, que pode ser utilizado para o reconhecimento de fala, ao identificar as sequências de fonemas de forma individual, em que cada estado acaba produzindo resultados observáveis com uma certa probabilidade de refletir o significado e a classificação correta dos termos. (KHURANA et al., 2017).

As abordagens estatísticas estão sendo utilizadas principalmente para os processos de reconhecimento de voz, traduções, aprendizado de novas gramáticas, aprendizados de novos elementos léxicos, entre outros.

A terceira abordagem utilizada é a conexionista. Essa abordagem é semelhante à estatística, pois desenvolvem modelos generalizados a partir dos fenômenos linguísticos que são obtidos em conjuntos de informações. Destaca-se que esse modelo combina a aprendizagem estatística com diversas teorias de representação, porém os sistemas conexionistas tornam mais difícil observar os modelos linguísticos, pois estes modelos são mais restritos do que aqueles que utilizam a abordagem estatística. (LIDDY, 2001).

Devi e Ponnusamy (2016, p. 193, tradução nossa) complementam tal questão, tratando das diferenças entre o modelo conexionista com os demais modelos, afirmando que:

> O que separa o conexionismo de outros métodos factuais é que os modelos conexionistas consolidam a aprendizagem factual com diferentes teorias de representação; assim, as representações conexionistas permitem a transformação, a inferência e a manipulação de fórmulas lógicas.

Um exemplo de modelo conexionista é chamado de modelo distribuído, que tem como princípio a ativação simultânea de diversas unidades, em que cada unidade participa de uma única representação conceitual. Tal modelo é mais utilizado nas tarefas de processamento de linguagem natural, por meio da análise sintática, estando focada mais na conversão do domínio, e na realização de associações para a definição sintática. (LIDDY, 2001).

Todas as abordagens apresentadas apresentam diferenças e são complementares de acordo com as necessidades do problema tratado no âmbito do processamento de linguagem

natural. Destaca-se que a evolução da área de PLN só foi possível pela existência dessas diversas formas de realizar as atividades e o processamento da linguagem.

Liddy (2001, p. 13, tradução nossa) demonstra, ainda, as relações entre as abordagens, relatando que:

> [...] as abordagens simbólica, estatística e conexionista exibiram características diferentes; assim, alguns problemas podem ser mais bem enfrentados com uma abordagem, enquanto outros problemas, por outra. Em alguns casos, para algumas tarefas específicas, uma abordagem pode ser adequada, enquanto em outros casos, as tarefas podem ficar tão complexas que pode não ser possível escolher uma única melhor abordagem.

Essa afirmação evidencia que é necessário utilizar diversas abordagens para ter resultados satisfatórios, ainda mais quando as tarefas que necessitam ser tratadas são demasiadamente complexas.

#### 4.3.5.2 Estratégias de processamento de linguagem natural

Relacionado às abordagens apresentadas, o processo realizado de PLN exige que algumas estratégias sejam adotadas. A seguir apresentam-se estratégias que estão vinculados às abordagens simbólicas, estatísticas e conexionistas; destacam-se as estratégias que aplicam o conhecimento linguístico: etiquetagem de texto, normalização de variações linguísticas e eliminação de *stopwords*, além da aplicação de métodos estatísticos.

A primeira técnica é a etiquetagem de texto, em que são inseridas etiquetas gramaticais do texto, sendo uma das primeiras estratégias realizadas nos sistemas de PLN. O etiquetador morfológico insere informações de categorias morfológicas, tais como substantivos e adjetivos. Por sua vez, o etiquetador sintático insere informações de funções sintáticas, como por exemplo, qual é o sujeito e o objeto da frase. Por fim, o etiquetador semântico introduz informações quanto ao significado dos termos, podendo inserir informações sobre quem é o agente e o seu estado. (GONZALEZ; LIMA, 2003).

Outra técnica está na normalização de variações linguísticas, que busca facilitar a compreensão dos termos normalizando nas esferas morfológica, sintática e léxico-semântica. A normalização morfológica realiza alguns processos como o *stemming* (reduz as palavras ao seu radical, retirando prefixo e sufixo) e a redução à forma canônica (geralmente, converte o verbo para o infinitivo e os substantivos para a forma masculina do singular). A outra normalização é a sintática, que busca normalizar as frases, de modo que frases que tenham um mesmo sentido, mas que estejam estruturadas de forma diferente, são representadas do mesmo

modo. E a normalização léxico-semântica visa agrupar palavras com similaridade semântica, encontrando um termo que signifique diversos outros. (GONZALEZ; LIMA, 2003).

A última técnica referente à aplicação de conhecimentos linguísticos está na eliminação de *stopwords*. Essa estratégia é utilizada em diversos contextos, e é uma importante técnica para os sistemas de processamento de linguagem natural. As *stopwords* são termos que, em geral, não contribuem para fornecer significado para uma determinada sentença, sendo, geralmente, artigos, conectivos e preposições. Destaca-se que se deve ter cuidado nesse processo, para não perder algumas importantes informações para a compreensão da estrutura do texto. (GONZALEZ; LIMA, 2003).

As três técnicas apresentadas são importantes para compreender e padronizar as estruturas textuais. Isso é necessário, pois os textos e a fala apresentam muitos termos distintos, que podem ser substituídos por elementos mais expressivos. Destaca-se ainda que o processo de etiquetagem é necessário para inserir informações que possam ser compreendidas para o computador. No geral, as técnicas apresentadas buscam aproximar a linguagem natural, que possui muitas variações, para algo que pode ser mais bem compreendido pelo computador.

A próxima estratégia trata da aplicação de métodos estatísticos. Esses métodos estão contribuindo significativamente para a evolução da área de PLN, em especial, devido à forte relação entre a área da estatística com a área de Inteligência Artificial. Destacam-se dois métodos importantes que têm relação com a área de PLN, a lei de Zipf e do gráfico de Luhn.

A primeira, lei de Zipf (ZIPF, 2016), foi proposta em 1949, e trata em suma de um método que por meio de uma base matemática e linguística, faz uma análise da frequência das palavras em um texto, considerando ainda o modo como tais termos estão distribuídos. Assim, com esse método, é possível ordenar quais as palavras que são mais presentes em um texto analisado. (CASSETARI, et al., 2015).

Nesse contexto, Gonzalez e Lima (2003, p. 13) explicam com mais detalhes o funcionamento desse método:

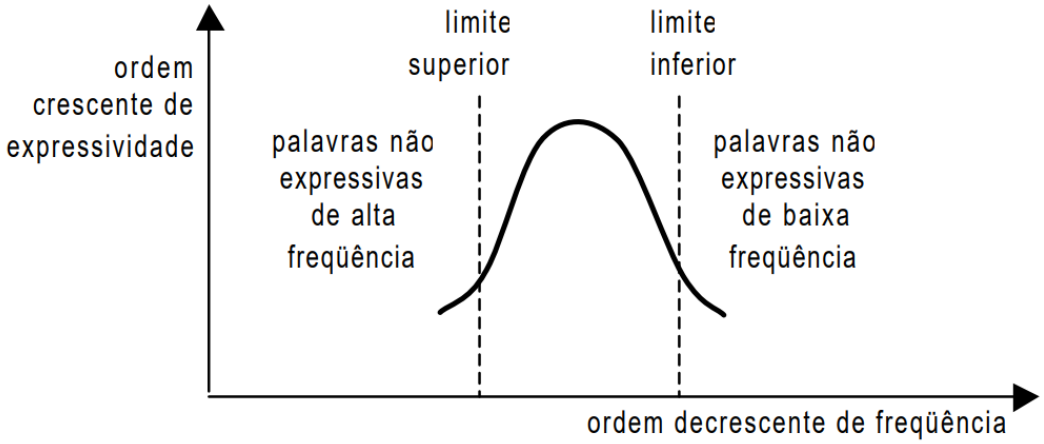
> Esta lei define que, tomando um determinado texto, o produto $\log(\mathbf{f_t})$ x $\mathbf{k_t}$ é aproximadamente constante, onde $\mathbf{f_t}$ é o número de vezes que o termo t ocorre no texto e $\mathbf{k_t}$ é a posição deste termo em uma relação de todos os termos daquele texto, ordenados pela frequência de ocorrência.

No geral, a lei fornece um modo para identificar os termos mais presentes, além de considerar a posição relativa desse termo, comparado aos demais. Utilizam-se, para isso, fórmulas matemáticas e estatísticas para posicionar e realizar as comparações entre os termos, definindo assim as análises.

Outro método bastante utilizado é o gráfico de Luhn. Ele propôs esse método em 1958, em que a frequência de aparição de um termo em um texto é uma importante medida sobre a sua expressividade. Essa técnica se baseia no princípio de que um autor, ao escrever um texto, para construir sua linha de raciocínio, irá utilizar uma grande quantidade de palavras que podem ser consideradas pouco expressivas; normalmente são artigos, preposições, e outras palavras sem muita relevância. Na outra ponta, as palavras com pouca frequência, também são consideradas pouco expressivas, uma vez que o autor não as utiliza com muita frequência, demonstrando que esses não são os principais termos do texto. Assim, as palavras que têm frequência intermediária são aquelas consideradas mais importantes no texto. (GONZALEZ; LIMA, 2003).

A figura 24 demonstra o gráfico que embasa o princípio desse método.

**Figura 24** - Gráfico de Luhn




Fonte: Gonzalez e Lima (2003, p. 13).

O gráfico demonstra a relação entre a frequência e a expressividade de um texto, demonstrando que os termos mais relevantes estão nessa faixa intermediária.

No contexto do PLN, as duas técnicas apresentadas podem ser importantes para definir quais são os principais termos de um texto. Esse procedimento é fundamental, quando há um grande conjunto de textos e informações e é necessário definir quais são os principais termos, e as palavras-chave, por exemplo. Destaca-se que essa técnica, normalmente, é utilizada apoiando outras, para definir os principais elementos, realizar classificações, entre outros usos.

O uso da estatística continua sendo muito utilizado no âmbito do processamento de linguagem natural, e apoiando técnicas de aprendizado de máquinas. Em um contexto de muitos

dados e muitos textos, é fundamental apoiar-se nesses métodos estatísticos para ter resultados que são satisfatórios, e capazes de realizar análises com grande escala.

Gonzalez e Lima (2003, p. 14) ainda apontam que: “Os métodos estatísticos podem ser utilizados para auxiliar o PLN em diversas situações. Eles têm sido utilizados na etiquetagem gramatical, na resolução de ambiguidade e na aquisição de conhecimento lexical [...], entre outras aplicações.”, revelando algumas das principais utilizações das técnicas de estatísticas no âmbito do PLN.

As estratégias relatadas, junto às diversas abordagens que foram destacadas, demonstram as possibilidades e limitações existentes no âmbito do PLN. A área do PLN contém diversas aplicações, como relatadas anteriormente e neste trabalho, em especial, será tratado o contexto do *Question Answering*. Nesse sentido, ao apresentar as abordagens e os diversos níveis de processamento de linguagem natural, é possível compreender melhor como essa área pode ser utilizada para favorecer o processo de recuperação da informação. Assim, todos os conceitos, destacando os níveis de PLN, bem como abordagens e estratégias, guiarão o processo de compreensão e de definição de como a recuperação da informação será tratada, tendo ênfase no *Question Answering*. Apresentam-se a seguir os principais conceitos envolvendo o *Question Answering*.

## 4.4 *QUESTION ANSWERING*

Uma importante corrente do desenvolvimento de soluções de recuperação da informação começou a utilizar os conceitos de processamento de linguagem natural (PLN) para desenvolver ferramentas e tecnologias para a realização de buscas. Assim, essas ferramentas buscam analisar de forma automática a linguagem natural humana utilizando algoritmos computacionais para encontrar o que o usuário deseja. A principal técnica expoente desse cenário é o *Question Answering* (QA), que são sistemas que têm como objetivo fornecer informações precisas e diretas respondendo a uma pergunta construída por um usuário.

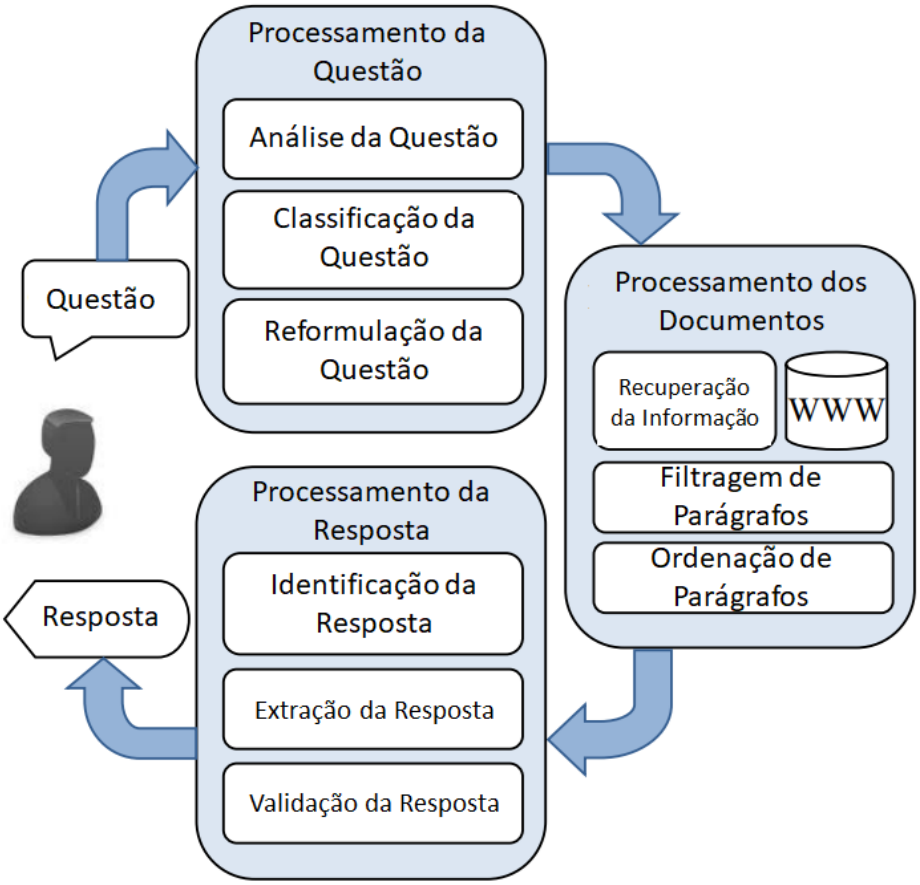
As técnicas de QA utilizam algoritmos avançados que permitem uma compreensão da linguagem natural, em que um usuário expressa por meio de uma pergunta em linguagem natural uma informação que ele deseja (ALMANSA, 2016). Um exemplo seria o usuário questionar: “Quais livros que tratam de Web Semântica estão disponíveis para empréstimo em uma biblioteca?”, necessitando que um sistema de QA compreenda esta questão, respondendo

exatamente o que o usuário deseja, fornecendo, por exemplo, uma lista com os livros disponíveis sobre essa temática.

Allam e Haggag (2012) afirmam que os estudos de QA abrangem pesquisas dentro de recuperação da informação, extração da informação e PLN. Os autores complementam relatando que: "[...] o principal objetivo de todos os sistemas de QA é recuperar respostas a perguntas em vez de documentos completos ou os melhores trechos desses documentos, como a maioria dos sistemas de recuperação de informação faz atualmente." (ALLAM; HAGGAG, 2012, p. 211, tradução nossa).

Um esquema que retrata os passos de um sistema de QA foi elaborado por Allam e Haggac (2012) e divide em cinco partes um sistema de QA: pergunta, processamento da pergunta, processamento do documento, processamento da resposta e resposta. A figura 25 apresenta a arquitetura clássica de um sistema de *Question Answering*, contendo as cinco partes destacadas pelos autores.

**Figura 25 -** Arquitetura do sistema de *Question Answering*




Fonte: Adaptado de Allam e Haggac (2012, p. 212)

A arquitetura apresentada na figura 25 consiste em 10 etapas, que serão apresentadas a seguir.

1. A pergunta consiste na fase inicial do processo, em que o usuário insere uma pergunta no sistema;
2. Na sequência é feito o processamento dessa pergunta, realizando uma análise dela, identificando o foco da pergunta, visando a aprimorar a qualidade do sistema;
3. Em seguida, é realizado o processo de classificação, para identificar o tipo de questão, além do tipo de resposta que deverá ser fornecida;
4. É realizada também uma reformulação da questão, expandindo-a, de modo a torná-la aderente para a recuperação da informação;
5. Na fase de processamento do documento, primeiramente é realizada a recuperação da informação relevante no contexto da pergunta realizada; esse processo acontece principalmente pela identificação de palavras-chave identificadas na questão;
6. Na sequência, é realizada uma filtragem da parte dos documentos encontrados, além de realizar um processo que encurta esses trechos;
7. Em seguida, esses parágrafos encontrados são ordenados, de modo que possam ser passados para o módulo de processamento de resposta;
8. Embasado no tipo de questão que foi realizada e no tipo de resposta esperada, são identificados os trechos que possam ser utilizados, chamados de respostas candidatas;
9. Após isso, utiliza-se um conjunto determinado de heurísticas para determinar e extrair apenas as palavras ou frases relevantes, capazes de responder à pergunta realizada;
10. E, finalmente, a resposta é validada, para ser apresentada ao usuário.

A explanação dessa proposta demonstra os principais processos de um sistema de QA, apontando os passos que um sistema de recuperação da informação desse tipo realiza. Há diversos sistemas de QA que realizam suas buscas dentro da Web ou em bases de periódicos científicos, auxiliando a pesquisadores e a outros usuários em suas buscas.

O processo detalhado será explanado na sequência, conforme relatado por Allam e Haggag (2012).

O primeiro módulo, parte desse sistema, é o de processamento da questão. A principal função desse módulo está na representação de uma questão em uma representação de

informação que possa ser processável. Em suma, esse módulo funciona em três etapas: análise da questão para encontrar a informação principal que o usuário deseja; classificação da pergunta para determinar o que o usuário espera como resposta; e reformulação da questão para converter o que o usuário necessita em algo que possa ser recuperado.

A análise da questão, também conhecida como foco da questão, busca identificar qual é a principal informação que o usuário está buscando naquela frase. O problema desse processo é que questões que iniciam com os termos “o quê” ou “para quê”, podem ser bastante ambíguas, então é fundamental analisar qual é o foco dessa questão. Por exemplo: uma pergunta como “Qual é o estado mais populoso do Brasil”, tem o foco no estado mais populoso; assim, tendo isso identificado, é mais fácil encontrar a resposta para tal questão.

A segunda etapa trata da classificação da pergunta e da resposta. A classificação da pergunta é necessária para entender qual é o tipo de informação que o usuário está desejando ao realizar aquela pergunta. Em suma, a pergunta é classificada dentro de um escopo tal como: perguntas do tipo: “o quê”, “por quê”, “quem”, “como”, “quando”, “onde”, entre outras. Já a classificação da resposta é uma consequência da classificação da pergunta, em que, de acordo com o que o usuário está desejando na pergunta, será dada uma resposta, que pode estar estruturada de diversas formas.

O próximo processo é a reformulação da pergunta, em que as palavras-chave, o foco e o tipo da pergunta são utilizados para criar os termos de recuperação da informação. Para isso, são aplicadas técnicas de PLN, para retirar *stopwords*, encontrar sinônimos, entre outras, além de utilizar taxonomias e tesauros para expandir a busca. Assim, com isso, tem-se uma questão reformulada, utilizando os princípios da recuperação da informação.

O segundo módulo é o de processamento do documento. Esse módulo irá buscar a questão reformulada em uma base de documentos, recuperando uma lista de documentos relevantes. Ainda nesse módulo ocorre o processo de filtragem e ordenação dos parágrafos, que serão trabalhados nos módulos seguintes. Assim, esse módulo está dividido em três partes: recuperação da informação, filtragem de parágrafo e ordenação de parágrafo.

A recuperação da informação acontecerá em diversos domínios, de acordo com a abrangência e o objetivo de cada sistema de *Question Answering*. Em suma, o processo é bastante semelhante, mas ao considerar as variáveis de precisão e recuperação, o sistema de recuperação da informação de um *Question Answering* está mais preocupado com aquilo que foi recuperado, do que com sua precisão; isso ocorre, pois em momento posterior será realizado outro processo que irá tratar aquilo que foi encontrado.

Após recuperar os documentos, realiza-se um processo de filtragem dos parágrafos. O princípio desse processo está em identificar, dentro dos documentos relevantes, os principais parágrafos que podem conter a resposta para a pergunta do usuário. Assim, parte-se da ideia de que a resposta está concentrada em partes do texto, e não espalhada em todo ele.

Em seguida, com a filtragem dos parágrafos ocorre o processo de ordenação. Esse processo busca ordenar os parágrafos na ordem de que a resposta mais viável está contida ali dentro. Para isso, aplica-se o algoritmo radix, que utiliza três pontuações para ordenar os parágrafos. 1) Pontuação da sequência de palavras: os termos da pergunta são reconhecidos na mesma sequência em um parágrafo; 2) Pontuação de distância: o quão distante estão os termos-chave no parágrafo; 3) Pontuação de palavra-chave ausente: a quantidade de palavras-chave que não estão presentes no parágrafo. Assim, a partir desses critérios, são ordenados os parágrafos.

O terceiro módulo é o módulo de processamento de resposta. Esse módulo deverá identificar, extrair e validar as respostas que foram encontradas, para, ao final, apresentar para os usuários. Tem-se assim esses três processos: análise, extração e validação.

O processo de identificação ocorre utilizando um conjunto de heurísticas, que visam a identificar as respostas candidatas. Para isso, é necessário reconhecer as entidades, como pessoas e objetos, para possibilitar encontrar tais respostas dentro dos parágrafos. O final desse processo cria, então, as respostas candidatas.

Em seguida ocorre a extração da resposta. Esse processo irá, a partir das respostas identificadas, extrair os resultados que o usuário está buscando em sua pergunta, utilizando uma série de heurísticas. Para isso, utilizam-se as palavras-chave, a distância entre elas, e outras métricas, para buscar encontrar a resposta que seja certeira. Caso isso não ocorra, apresenta-se o parágrafo que mais se aproxima disso. Vale destacar, ainda, que os sistemas de *Question Answering* devem apresentar uma única resposta como correta.

Por fim, ocorre o processo de validação da resposta. Esse processo busca validar se a extração encontrou uma resposta que de fato é coerente e satisfaz as necessidades do usuário ao realizar aquela pergunta. Para isso, utilizam-se recursos léxicos ou realiza-se outra busca da Web, buscando validar se o resultado é coerente, de acordo com a quantidade de resultados obtidos. Destaca-se que é importante esse processo para que o sistema seja confiável e apresente bons resultados.

Complementarmente, ressalta-se que, com a evolução dos sistemas de *Question Answering*, algumas propostas buscaram inserir mecanismos semânticos a esses sistemas, visando aprimorar as respostas realizadas. Pesquisas começaram a utilizar o *Linked Data* como

fonte de informação, na busca de ter respostas mais precisas. Nesse sentido, Unger et al. (2012, tradução nossa) relatam que essas pesquisas tinham como base utilizar a estrutura de RDF para o *Question Answering*, em que esses sistemas: "[...] mapeavam uma questão em linguagem natural para uma tripla RDF."

Diante disso, o modelo criado que será apresentado na seção 5 demonstra a aproximação entre processamento de linguagem natural, Web Semântica, *Question Answering* e *Linked Data*.

Um aspecto central para a criação desse modelo está na área de *machine learning*, parte da Inteligência Artificial, que será explorada a seguir.

## 4.5 MACHINE LEARNING – APRENDIZAGEM DE MÁQUINAS

A área de aprendizagem dentro da Inteligência Artificial é um dos campos mais estudados e aplicados atualmente. Existem diversos sistemas que estão inserindo tais tecnologias no dia a dia das pessoas, criando uma série de facilidades para diversas funções e se tornando um elemento indispensável para as organizações. Exemplos como assistentes virtuais de bancos e operadores de telemarketing virtual utilizam outros elementos da Inteligência Artificial, como o processamento de linguagem natural, mas tem no aprendizado de máquina a sua principal função: a capacidade de "aprender".

Uma definição de *machine learning* é dada por Jordan e Mitchell (2015, p. 255, tradução nossa):

> O aprendizado de máquina é uma disciplina focada em duas questões inter-relacionadas: Como construir sistemas de computador que melhoram automaticamente com a experiência? e Quais são as leis fundamentais da teoria estatística da informação computacional que governam todos os sistemas de aprendizagem, incluindo computadores, seres humanos e organizações? O estudo do aprendizado de máquina é importante tanto para abordar essas questões científicas e de engenharia fundamentais, quanto para o software de computador altamente prático que ele produziu e utilizou em vários aplicativos.

A definição apresentada coloca dois importantes elementos para apontar o que é *machine learning*. O primeiro trata do aspecto da máquina conseguir evoluir a partir das experiências e do uso que vai ocorrendo. Esse primeiro ponto é, atualmente, o mais conhecido das aplicações de aprendizagem de máquinas, e demonstra um dos princípios dessa área, que está em, a partir do uso, fazer acontecer o aprimoramento da aplicação.

O segundo aspecto dessa definição está na questão estatística, em que o uso dessa ciência contribui para compreender as relações entre computadores, seres humanos e organizações. O uso de estatística é essencial na área de *machine learning*, pois o computador,

por ter todo o seu funcionamento baseado em leis matemáticas, tem que usar da estatística para conseguir aprender e traçar possibilidades visando dar uma resposta satisfatória a um determinado problema humano.

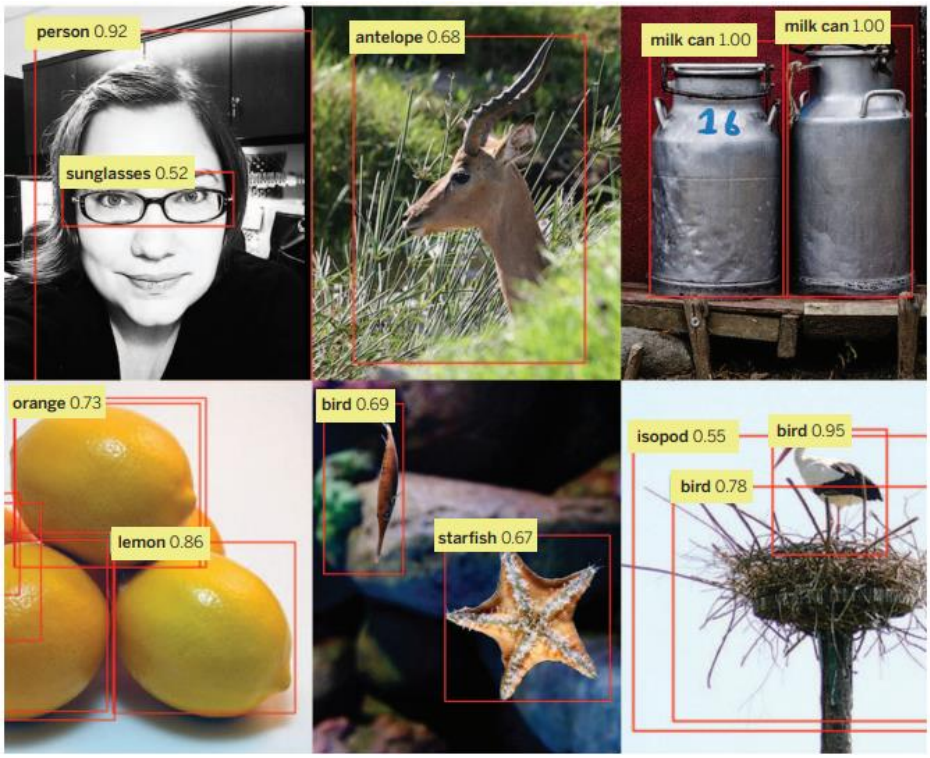
Um terceiro aspecto trazido pela definição apresentada contrapõe dois importantes elementos dessa área, que diz respeito ao objetivo de se utilizar e de se aplicar *machine learning*. O primeiro trata da aplicação em questões científicas e de engenharia, que utilizam os algoritmos para fazer avançar a ciência e conseguir grandes inovações. O outro aspecto é o caráter prático, que vem sendo aplicado em soluções do cotidiano das pessoas, que pode ser visto em soluções utilizadas por milhões de usuários, como assistentes virtuais de dispositivos computacionais.

Nesse sentido, relacionado a tais pontos, se encontram os algoritmos de *machine learning*, que são capazes de executar os pontos destacados anteriormente. Um algoritmo de *machine learning*, conceitualmente, funciona como uma pesquisa em um grande espaço com programas candidatos, e levando em consideração as experiências anteriores e os treinamentos realizados, busca encontrar o programa com melhor desempenho. Esses programas candidatos são, por exemplo, árvores de decisão ou funções matemáticas. (JORDAN; MITCHELL, 2015).

Há diversas aplicações, nos mais diversos contextos, que estão utilizando *machine learning*. Destaca-se o uso em veículos autônomos, em projetos de processamento de linguagem natural, buscadores e em aplicações de visão computacional.

A figura 26 apresenta um exemplo de aplicação de *machine learning*, em que o algoritmo busca identificar quais são os objetos que estão aparecendo na imagem, sendo um projeto na área de visão computacional.

**Figura 26** – Exemplo de aplicação de *machine learning*




Fonte: Adaptado de Jordan e Mitchell (2015, p. 256)

Na figura 26, verifica-se a busca por identificar quais são os objetos que se encontram na imagem. Esse processo acontece a partir do reconhecimento de características das imagens, e o uso de *machine learning* ocorre por meio de treinamentos, em que, anteriormente ao processo de aplicação, o algoritmo é treinado com várias imagens em que é explicitado para a máquina qual é o objeto que está presente em cada imagem. Essa fase, chamada de treinamento, é essencial para que o algoritmo possa depois ser validado e utilizado na prática. Ao final, com o algoritmo treinado, pode ser realizada a identificação dos objetos, em que se aponta qual é o objeto e verifica-se a probabilidade daquilo ser aquele objeto mesmo. Na figura, percebe-se que junto ao nome, há uma porcentagem, e em alguns casos existem erros, o que é natural, pois a depender do ângulo e das características da imagem, o algoritmo pode errar.

Assim, a visão computacional que traz essa interpretação dos dados a partir das características visuais das imagens (BARROW; TENEMBAUM, 1981) evoluiu significativamente com a aplicação do *machine learning*, sendo uma de suas principais aplicações atualmente.

Complementarmente, outro autor aprofunda tal questão ao tratar da abordagem probabilística da aprendizagem de máquinas:

> As abordagens probabilísticas do aprendizado de máquina e da inteligência são uma área de pesquisa muito ativa, com amplo impacto, além dos

> problemas tradicionais de reconhecimento de padrões. Conforme destacamos, esses problemas incluem compactação de dados, otimização, tomada de decisão, descoberta e interpretação de modelos científicos e personalização. (GHAHRAMANI, 2015, p. 11, tradução nossa).

A abordagem apresentada é importante para demonstrar como a área de *machine learning* é amplamente utilizada no contexto de análise de dados em cenários de *Big Data*. A grande quantidade de dados seria impossível de ser amplamente analisada caso não fizesse uso de técnicas de aprendizagem de máquinas, em especial da abordagem probabilística.

Um importante aspecto acerca dos algoritmos de *machine learning* é o processo de como que ele aprende e os modos como isso pode acontecer. O primeiro passo é o treinamento, em que se utiliza um conjunto de dados pré-definidos, para que a máquina consiga identificar quais são as variáveis para que aquele determinado resultado seja alcançado. Vale destacar que se busca definir um conjunto de dados de treinamento que seja relevante para aquele cenário, pois o algoritmo irá utilizar tais dados e variáveis para encontrar o padrão que define o resultado. (RIBEIRO; FRAZÃO; SA, 2018).

Além disso, o modo como a previsão é realizada pode ser categorizado em três tipos: aprendizado supervisionado, aprendizado não-supervisionado e semi-supervisionado. (LAMPROPOULOS; TSIHRINTZIS, 2015).

O primeiro tipo, aprendizado supervisionado, utiliza dados para treinamento, cujo resultado é conhecido e explicitado para o algoritmo. Assim, o algoritmo conhece a solução e a partir dele e dos dados definirá quais são os aspectos que devem ser considerados para classificar algo em uma categoria. No segundo tipo, aprendizado não-supervisionado, não há o resultado ou a solução desejada, utilizando padrões estatísticos nos conjuntos de dados de treino. E no terceiro tipo, aprendizagem semi-supervisionada, os dados são parcialmente rotulados, em que parte das informações está com soluções, mas outra parte, com outros tipos de dados, não.

No geral, todos os algoritmos de *machine learning* tem uma base de treinamento, que permite que o algoritmo, ao entrar em execução, possa se utilizar dela, para gerar os resultados das análises. Assim, *o machine learning* tem sempre que consultar um conjunto de dados, e quanto melhor e com mais tempo de uso, o algoritmo tende a ser aprimorado e ter resultados mais efetivos.

A seguir apresenta-se a seção que discute a aproximação entre a recuperação da informação, a Inteligência Artificial e a Web Semântica.

# 5 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO UTILIZANDO INTELIGÊNCIA ARTIFICIAL E WEB SEMÂNTICA

Esta seção desenvolve o presente trabalho visando a contribuir para a área de recuperação da informação e Ciência da Informação, posicionando os campos da Inteligência Artificial e Web Semântica nesse contexto. A relação interdisciplinar entre Ciência da Computação e Ciência da Informação é parte de todo o desenvolvimento do trabalho, aperfeiçoando e criando perspectivas para o processo de recuperação da informação.

Ao final da seção, realiza-se uma prova de conceito do trabalho, para demonstrar a validade do modelo proposto.

## 5.1 INTELIGÊNCIA ARTIFICIAL E A WEB SEMÂNTICA

As ferramentas da Web Semântica têm sido utilizadas para aprimorar o modo como as informações são compreendidas em diversos contextos, como recuperação, organização e tratamento da informação, além de seu uso em processamento e análise de dados em áreas vinculadas à Ciência de dados.

Em especial na recuperação da informação, as ferramentas da Web Semântica estão contribuindo para transformar a recuperação sintática de documentos, em uma recuperação que considera o significado e o contexto no qual os termos se encontram. Destacam-se tecnologias como as ontologias, com OWL, além do RDF e SKOS, para criar vocabulário e estruturar os dados, e o SPARQL para favorecer a recuperação de dados.

Apesar das claras contribuições da Web Semântica no aprimoramento dos processos computacionais, entre eles o de recuperação, tais ferramentas possuem um certo limite no sentido de efetivamente levar à compreensão semântica do contexto de determinados conjuntos de dados. Isso porque o processo, no que tange ao uso das ferramentas da Web Semântica, exige que profissionais participem da elaboração dos artefatos, como, por exemplo, na elaboração de ontologia, ou na publicação de dados em formatos de *Linked Data*.

Outro aspecto a ser considerado está no uso cada vez mais efetivo de Inteligência Artificial em diversas aplicações, bem como a sua rápida evolução, que tem trazido importantes contribuições, especialmente em contextos em que há a oportunidade de automatizar os processos. Uma subárea da Inteligência Artificial é o campo do processamento de linguagem natural, que tem sido bastante utilizado para a compreensão de termos e frases diversas.

O processamento de linguagem natural possui uma série de aplicações, como a classificação de termos, a compreensão do sentido, a identificação de relações, entre outras.

Esses tipos de aplicações de PLN estão evoluindo significativamente, e são realizados por meio de algoritmos de aprendizagem de máquinas e consulta a diversas bases de dados, obtendo resultados bastante expressivos, mas que são limitados àquilo que já existe nessas bases. Além disso, não existe um nível de semântica elevado quanto às propriedades e aos tipos de relacionamentos que possam existir entre os termos.

Isso é apontado por diversos autores, como Liddy (2001), que afirma que os níveis de processamento de linguagem natural que normalmente são aplicados são os mais baixos, como o da fonologia, da morfologia e do léxico, pois os níveis mais elevados, como sintático, semântico, discursivo e pragmático são regidos apenas pela regularidade, o que torna o processo de PLN muito mais complexo. A autora complementa que há poucos sistemas que estão utilizando efetivamente os níveis mais elevados de PLN.

A partir disso, verifica-se que as ferramentas da Web Semântica são capazes de fornecer elementos com um elevado nível de semântica formal, mas isso sempre depende de considerações humanas para um aprimoramento e para que seja possibilitada uma visão mais geral desses elementos, que dependem de ontologias de temas específicos. Esse cenário é explicitado por Santarem Segundo e Coneglian (2016), que relatam como as ontologias agregam o nível de semântica e possibilitam aumentar o nível de inferência, porém os autores enfatizam que as ontologias necessitam de ferramentais para apoiá-las e extrair informações que de fato possam contribuir para encontrar conhecimento nas bases de dados.

Por outro lado, o processamento de linguagem natural se destaca, pois, ao usar algoritmos de aprendizagem de máquina, consegue aprimorar a compreensão de um cenário, utilizando informações que foram utilizadas para o treinamento desses algoritmos. No entanto, como relatado, o PLN não fornece o nível de semântica formal que as ferramentas da Web Semântica fornecem.

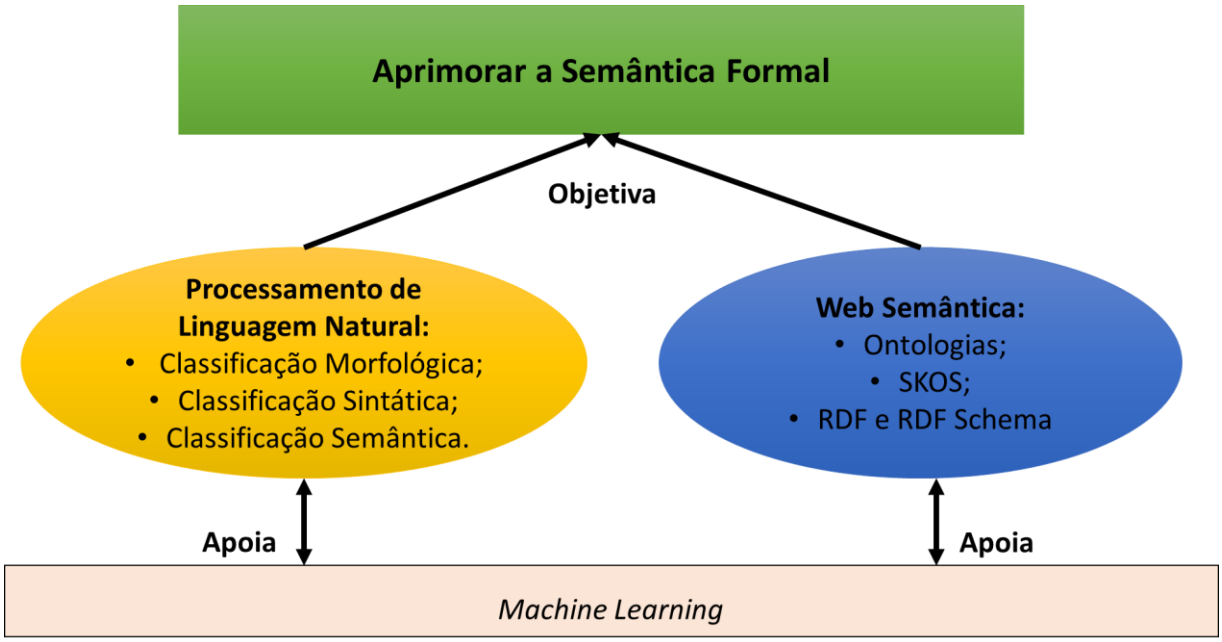
Esse cenário demonstra que a união das ferramentas e dos conceitos da Web Semântica e dos algoritmos e técnicas de processamento de linguagem natural é capaz de aprofundar significativamente a compreensão dos termos e aprimorar, assim, o processo de recuperação da informação.

Adicionalmente, um outro campo da Inteligência Artificial pode aprimorar ainda mais esse processo, a aplicação de *machine learning*. A inserção de aprendizagem de máquinas, tanto no PLN quanto para apoiar as ferramentas da Web Semântica, pode favorecer a compreensão e a interpretação do significado dos termos e das frases de forma a que sejam melhoradas. Isso porque as técnicas de *machine learning* podem utilizar o histórico e os treinamentos realizados para apoiar e definir palavras-chave e os tópicos que os termos possuem. Dessa forma, um

termo que está classificado em uma ontologia, pode ser comparado ou ter a sua compreensão aperfeiçoada a partir de uma análise de tópicos utilizando o algoritmo de *machine learning*.

A figura 27 estabelece uma relação entre a Web Semântica, o processamento de linguagem natural e *machine learning* no aspecto da semântica formal.

**Figura 27** - Semântica formal




Fonte: elaborada pelo autor.

Na figura 27, verifica-se que a semântica formal é o objetivo a se atingir, tanto do processamento de linguagem natural, quanto da Web Semântica, que visam a obter um grau mais elevado dessa semântica. Ambos estão ligados ao *machine learning*, que aprimora as possibilidades de se atingir tal objetivo. Além disso, são apontados alguns dos instrumentos que irão apoiar tanto a Web Semântica, quanto o processamento de linguagem natural.

No processamento de linguagem natural foram destacados vários momentos. O primeiro, a classificação morfossintática, que se refere ao processo de identificação de substantivos, artigos, verbos, entre outros, para, em seguida, detectar sua função sintática na frase. Essa análise sintática é essencial para posteriormente se chegar a uma compreensão do significado dos termos.

O primeiro elemento é a classificação de termos, que permite realizar classificações de classes e tipos a que um determinado termo pode pertencer. Essa classificação é importante para, em um segundo momento, verificar quais termos estão vinculados a um outro, além de realizar a classificação de conceitos, o que permitirá, depois, aprofundar a compreensão de

textos. Essa classificação busca trazer uma série de informações de um termo ou conceito que são essenciais para se ter um panorama geral do que um termo significa, além de possibilitar estabelecer que termos estão relacionados, assim como características de determinado conceito. Esse momento é o mais próximo de um alto nível de semântica formal, uma vez que fornece uma quantidade grande de informações sobre os termos.

Todos esses processos do PLN podem utilizar *machine learning*, uma vez que, ao usar algoritmos como o *topic modelling*, é possível extrair mais informações e ter uma melhor análise morfossintática e semântica dos textos.

Com relação à Web Semântica, tem-se também três elementos. O primeiro é o RDF, que não permite a definição de um nível de semântica formal aprofundado. No entanto, o RDF é essencial para a modelagem dos dados, embora ele, por si só, não seja capaz de expressar muitas informações que indiquem o significado de um termo.

O SKOS, por outro lado, é um vocabulário reconhecido para a construção de tesauros, com uma ampla gama de relações, e pode ser um importante instrumento para descrever um cenário. Os tesauros são elementos que têm um bom nível de semântica formal; assim, o uso do SKOS pode ser um auxiliar importante para se alcançar uma boa compreensão de termos e de seu contexto.

O terceiro e último elemento são as ontologias, instrumentos da Web Semântica que mais se aproximam de um nível elevado de semântica formal. As ontologias são capazes de mostrar altos níveis de semântica, com diversas propriedades e relacionamentos. Destaca-se ainda que apenas ontologias são capazes de fornecer informações mais contextuais, e isso as torna fundamentais para que a compreensão dos termos seja mais efetiva.

O apoio de algoritmos de *machine learning* para a realização de inferências pode auxiliar significativamente os resultados obtidos com as ontologias. Isso porque os axiomas podem ser expandidos com outros dados e, assim, ter um resultado mais eficiente na análise e inferência.

Por fim, é importante ressaltar que o modelo que embasa esta tese está construído em cima dessa premissa, em que Web Semântica e PLN se complementam e são capazes de aprimorar os diversos processos e, em especial, no âmbito deste trabalho, a recuperação da informação.

## 5.2 RECUPERAÇÃO DA INFORMAÇÃO E PROCESSAMENTO DE LINGUAGEM NATURAL: CONCEITUAÇÃO DO MODELO

Esta subseção apresenta os conceitos e as reflexões que conduziram à construção do modelo. Nesta subseção ainda não será apresentado o modelo, que virá na seção subsequente, mas sim a concepção e a reflexão que levaram ao seu desenvolvimento.

As questões apresentadas, relacionadas ao modo como a Web Semântica pode estar em conjunto com o processamento de linguagem natural, suscitam a necessidade de se reverem as formas pelas quais a recuperação da informação é trabalhada.

Adicionalmente, busca-se demonstrar que, ao realizar essa aproximação das ferramentas da Web Semântica com PLN e Inteligência Artificial, se tem uma recuperação da informação mais eficiente, pois permitirá um nível de semântica formal que não é alcançado nas pesquisas atuais, permitindo uma contextualização mais clara tanto das necessidades informacionais dos usuários, quanto dos documentos e informações recuperadas. Assim, o processo de recuperação da informação passa a considerar aspectos semânticos, por duas vias: a primeira devido ao uso das ferramentas da Web Semântica, e a segunda, pela aplicação das técnicas de PLN, que, atualmente, permitem a compreensão do sentido dos termos, bem como de conceitos relacionados.

Outro aspecto que é central para que a recuperação da informação seja primordialmente semântica está na adoção de fontes de informação que tenham alta expressividade e utilizem as ferramentas da Web Semântica. Nesse sentido, a possibilidade de se adotar como fonte de informação bases de dados publicadas seguindo os princípios do *Linked Data* pode contribuir para aumentar o nível de semântica formal da recuperação da informação.

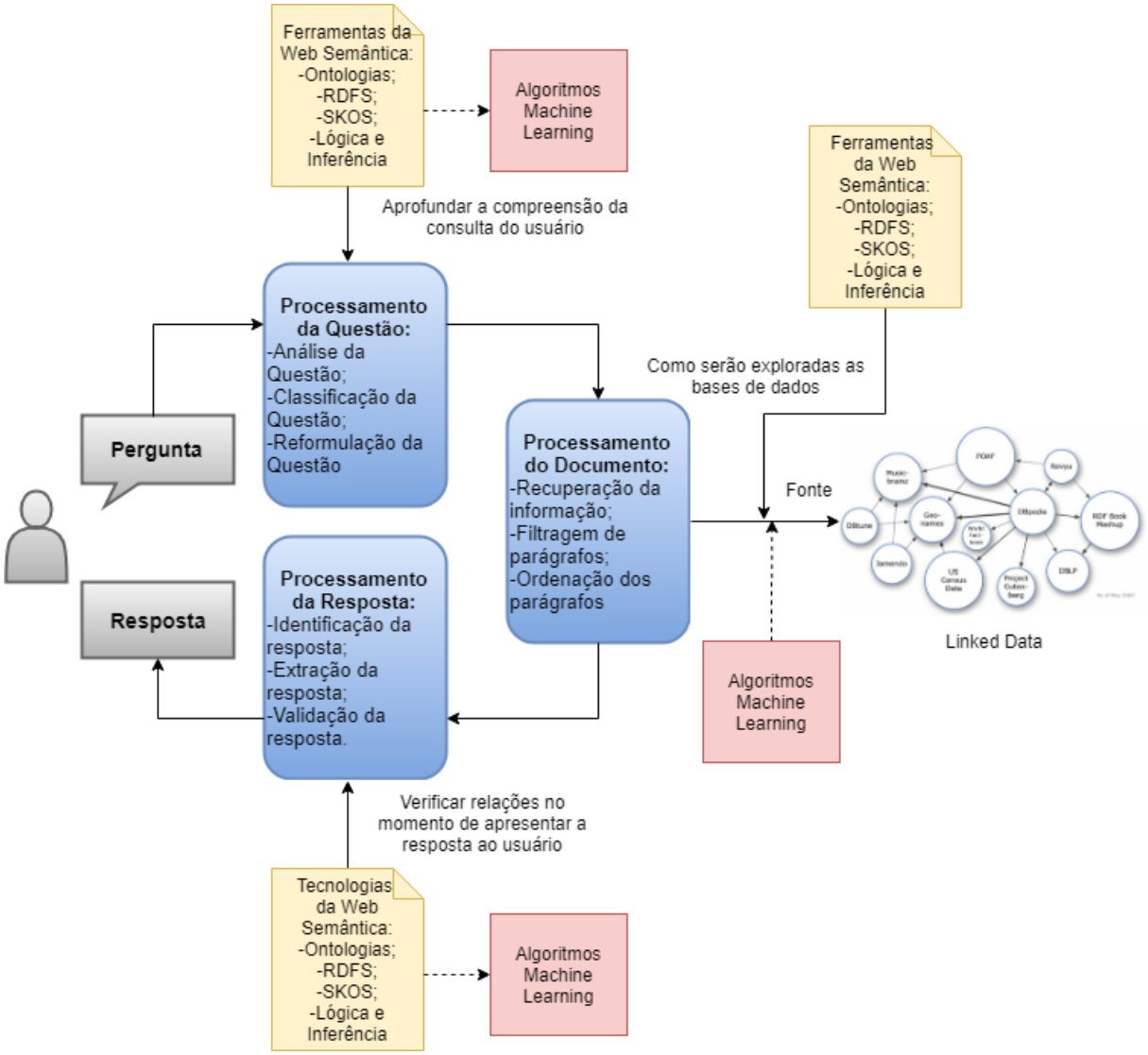
Os dados que seguem os princípios do *Linked Data* são publicados com propriedades e relações com um nível de semântica mais elevado, que fornece informações importantes para se ter uma recuperação da informação mais eficiente. Assim, quando comparados esses dados a bases de dados relacionais ou dados não estruturados disponíveis na Web, tem-se informações que são mais expressivas e que favorecem o processo de recuperação da informação.

Complementarmente, como dados em formato de *Linked Data* utilizam ferramentas da Web Semântica, como RDF e OWL, a relação com as ontologias que serão utilizadas para a compreensão do processo de busca se torna mais natural e, consequentemente, mais eficiente.

A figura 28 evidencia, a partir de um modelo clássico de *Question Answering*, como as ferramentas da Web Semântica e os dados de *Linked Data* podem ser inseridos no processo de aplicação do QA. Destaca-se que essa figura aponta apenas em quais momentos do QA serão inseridos elementos da Web Semântica, não pretendendo ser um modelo ou uma arquitetura. O

modelo base de *Question Answering* é inspirado na proposta feita por Allam e Haggag (2012), que foi apresentado anteriormente na seção de Inteligência Artificial.

**Figura 28** - Ferramentas da Web Semântica no *Question Answering*




Fonte: Adaptado de Allam e Haggac (2012, p. 212)

A figura 28 está dividida em três processamentos principais: questão, documento e resposta. Basicamente, após o usuário realizar a pergunta, o processo do *Question Answering* faz o tratamento dessa questão. Nessa etapa, as ferramentas da Web Semântica podem ser inseridas, ao aprofundar a compreensão que os mecanismos computacionais possuem dos termos e da pergunta realizada. Além disso, os algoritmos de *machine learning* apoiam o

processo de compreensão dos termos, ao utilizar algoritmos, como o *Topic Modelling* e *Latent Dirichlet allocation,* que extraem tópicos e se relacionam a outros termos.

Na sequência, realiza-se o processamento de documento, em que se deve encontrar os materiais que atendam a pergunta realizada pelo usuário. Nesse momento, pode-se utilizar como fonte de informação o *Linked Data*, fornecendo materiais com um nível de semântica mais elevado. Além disso, outras ferramentas da Web Semântica podem ser utilizadas para aprimorar a recuperação de documentos, inclusive nas fontes de *Linked Data.* Vale destacar que nessa etapa, os algoritmos de *machine learning*, que podem utilizar aprendizado supervisionado para, a partir de um banco de perguntas e respostas dadas anteriormente, auxiliar na expansão da busca, relacionam-se com o processo realizado pelas ferramentas da Web Semântica.

Por fim, no processamento da resposta, obtêm-se a resposta para a questão levantada pelo usuário. Esse processo novamente pode ser assistido e auxiliado pelas ferramentas da Web Semântica, ao considerar termos e conceitos da ontologia que possibilita o entendimento do contexto e do significado dos elementos que serão considerados para a definição da resposta apresentada ao usuário. Novamente, os algoritmos de *machine learning* podem auxiliar na validação dos resultados e definição dos melhores termos a serem apresentados aos usuários como resposta.

Vale destacar que, partiu-se do modelo de *Question Answering*, pois tal modelo já possui em suas bases a realização do processo de recuperação da informação com o uso do processamento de linguagem natural. No entanto, o modelo é expandido e alterado, uma vez que são inseridos elementos distintos como as próprias ferramentas da Web Semântica, a base de *Linked Data* e Inteligência Artificial.

A partir dos conceitos e das reflexões realizadas nesta subseção, definiu-se o modelo que será apresentado a seguir.

## 5.3 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO UTILIZANDO INTELIGÊNCIA ARTIFICIAL E PROCESSAMENTO DE LINGUAGEM NATURAL

Partindo da interseção base desta tese, que está na união entre os conceitos e as ferramentas da Web Semântica, os princípios da Inteligência Artificial, bem como o processamento de linguagem natural e a recuperação da informação, será apresentado e discutido o modelo proposto nesta seção.

A base do modelo está na proposta de recuperação da informação, inspirado nos conceitos do *Question Answering*, em que o usuário irá ter acesso e encontrar as informações

de que ele necessita por meio do uso de linguagem natural. Nesse sentido, o processo de recuperação da informação que tem como base o processamento de linguagem natural está estruturado no modo como a pergunta é compreendida, frente ao modo como a resposta será estruturada e apresentada ao usuário.

No entanto, trazendo o que foi proposto como tese, o modelo proposto busca inserir nesse processo de recuperação da informação, o uso mais efetivo de compreensão textual com a aplicação de técnicas de Inteligência Artificial. A ideia está em tornar a compreensão do que o usuário está querendo mais clara, bem como permitir que o tratamento e a recuperação das informações sejam mais eficientes. Isso será possível pela utilização de estruturas e técnicas da Inteligência Artificial que possibilite a classificação e a compreensão dos textos.

Cabe uma primeira ressalva ao uso do termo compreensão, no que tange à Inteligência Artificial. Quando se utiliza compreensão, nesse cenário, tem-se como perspectiva apresentar que a IA será utilizada visando a mostrar, em históricos recentes e estatísticos, o que aquela palavra pode significar, e com que termos aquele conceito está vinculado. Isso se faz necessário uma vez que é fundamental verificar, de uma forma automatizada, qual é o provável significado daquele termo; além disso, busca-se verificar quais conceitos podem estar vinculados a ele.

Adicionalmente, insere-se neste modelo a principal contribuição ao tratamento utilizando PLN, que está na aplicação dos conceitos e das ferramentas da Web Semântica, visando a tornar o processo efetivamente semântico, considerando o significado e o contexto no qual os termos se encontram. As ferramentas da Web Semântica e Inteligência Artificial serão adotadas buscando diminuir a lacuna (*gap*) existente entre o modo como as tecnologias computacionais tratam conceitos, em que o significado não é claro e não está definido efetivamente para o seu tratamento computacional.

Nesse contexto, ainda que a aplicação de técnicas de Inteligência Artificial possa dar uma visão geral, utilizando como premissas os usos anteriores de termos, existe a falta de uma compreensão de quais relacionamentos e significados os termos efetivamente possuem. A aplicação das ferramentas da Web Semântica juntamente com o uso de Inteligência Artificial e do processamento de linguagem natural podem tornar o processo mais efetivo, e se aproximando, de fato, de uma compreensão daquilo que uma pessoa está dizendo (ou escrevendo).

Por fim, uma característica central e determinante para tornar o processo de fato efetivo e eficiente está no tratamento e no uso de fontes informacionais que possuem um nível de semântica formal elevado. Nesse contexto, o uso de dados estruturados em formato de *Linked Data*, que utilizam modelos de dados e vocabulários expressivos e que estão publicados e

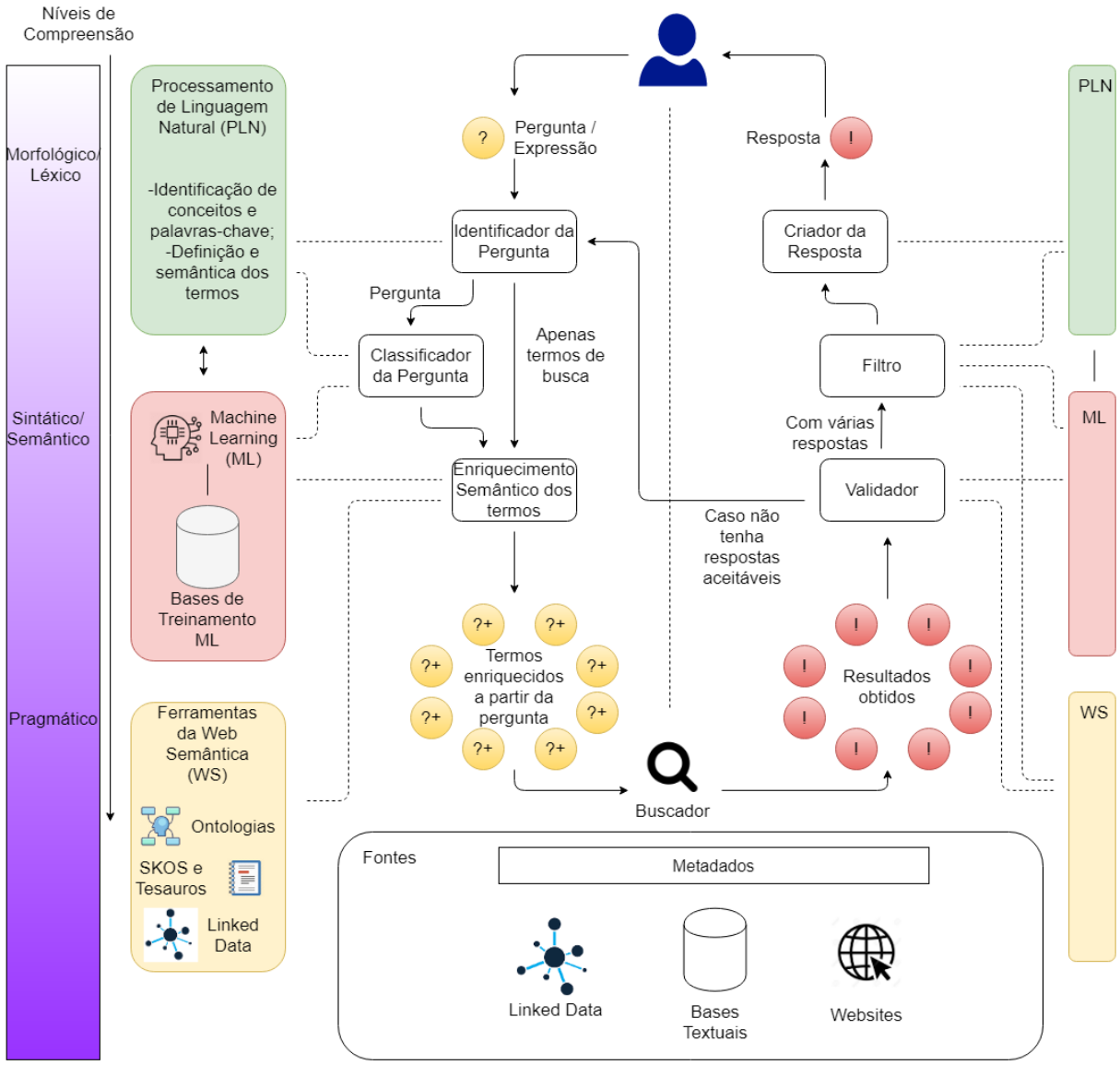
interligados a outros modelos de dados, torna a recuperação da informação mais precisa. Destaca-se que, ainda que a qualidade dos dados publicados no *Linked Data* não seja boa, o uso de dados publicados nesse formato permite que as ontologias que apoiam o modelo tenham uma maior aderência aos dados, permitindo que os resultados alcançados sejam melhores.

Isso ocorre pelo fato de que, ao utilizar dados que estão estruturados com base em ontologias, por exemplo, é possível ter uma compreensão mais clara do que as informações ali contidas significam. O princípio está em que, ao compreender o que uma pergunta significa, um sistema desenvolvido a partir do modelo proposto, teria como resultado grafos que traduziriam aquilo que o usuário deseja saber; com tais grafos, o acesso e a recuperação em fontes informacionais de *Linked Data* seria natural e, por consequência, facilitado. Sendo assim, a recuperação da informação consideraria uma série de aspectos semânticos para recuperar as informações que foram buscadas.

Vale destacar que o modelo proposto é aplicado apenas à ambientes informacionais digitais, pela necessidade da exploração do documento para a construção dos resultados e das respostas. Assim, não é possível aplicar tal modelo em ambientes analógicos, como em acervos de bibliotecas tradicionais.

Depois dessa breve explicação do modelo, apresenta-se a seguir, na figura 29, o modelo proposto nesta tese.

**Figura 29** - Modelo conceitual de recuperação da informação




Fonte: elaborado pelo autor.

O modelo apresentado na figura 29 está dividido em quatro partes principais: a primeira (superior central) trata da parte referente a todo o processo de recuperação da informação, e a segunda (inferior central) demonstra as fontes informacionais que serão utilizadas para possibilitar o processo de recuperação da informação. Nas duas laterais, encontram-se três importantes camadas: processamento de linguagem natural, *machine learning* e ferramentas da Web Semântica, responsáveis por apoiar o processo da recuperação da informação. E na lateral esquerda, está presente uma escala de nível de compreensão, utilizada para demonstrar o aprofundamento nos níveis linguísticos.

As três camadas das laterais serão chamadas de camadas de suporte, pois elas que viabilizam o modelo por meio de suas tecnologias e algoritmos. A primeira camada é a de processamento de linguagem natural, que é responsável por fazer o entendimento e identificação dos termos, definir as categorias de tais termos, apoiar a montagem da resposta e apoiar a classificação da pergunta.

Essa camada está fortemente vinculada à camada de *machine learning*, responsável por utilizar algoritmos, como o de ***topic modeling,*** como uma técnica de aprendizado não supervisionado, para auxiliar o processo de identificação dos termos e palavras-chave. Utilizam-se também técnicas de aprendizado supervisionado de **classificação**, que buscam classificar os conteúdos em categorias definidas. Com ambas as técnicas de *machine learning*, é possível considerar que, além disso, essa camada auxilia no processo de enriquecimento e validação, uma vez que as bases de treinamento do algoritmo irão promover uma constante evolução em comparação com outros textos e resultados.

A terceira camada é a de ferramentas da Web Semântica, focada, especialmente, no enriquecimento da pergunta e no apoio do filtro e da validação da resposta. Em suma, as ferramentas da Web Semântica participam das etapas que necessitam aumentar o nível de semântica formal, tornando os resultados mais aprimorados. As principais ferramentas da Web Semântica utilizadas são ontologias, modelos de dados em RDF e SKOS, regras e inferências com o SWRL e o SPARQL, para promover a recuperação de dados.

Quanto ao processo do modelo, que se encontra na parte central da figura 29, considere-se que esse processo está dividido em duas partes, pergunta e resposta, trazendo, assim, um conceito oriundo do *Question Answering*, em que o processo é tratado dessa forma: o usuário realiza a pergunta, o sistema trata e compreende essa pergunta, busca-se o resultado, e o sistema reage em formato de resposta, sendo essa a principal aproximação entre o QA e o modelo proposto.

Essa primeira parte do modelo é composta basicamente por sete módulos: identificador da pergunta, classificador da pergunta, enriquecimento semântico da pergunta, buscador, validador, filtro e criador da resposta, apresentados a seguir.

O primeiro módulo é o **identificador da pergunta** que tem como objetivo verificar se o que o usuário buscou é de fato uma pergunta, ou apenas termos de uma busca normal. Esse processo é importante, pois, a partir disso, o modelo irá realizar ou o tratamento daquela pergunta, direcionando para o módulo do classificador da pergunta, ou se for apenas termos de pergunta, já direciona diretamente para o enriquecimento semântico da pergunta.

No caso de pergunta, a próxima etapa é o **classificador da pergunta**, que irá receber a pergunta feita pelo usuário e irá realizar um processo de identificação e análise de, primeiramente, qual é o tipo de pergunta que o usuário está fazendo e, mais adiante, verificação de quais são os termos e pontos-chave que compõem essa pergunta. Para isso o classificador de perguntas irá se relacionar com os algoritmos de *machine learning*, para que, por meio do aprendizado supervisionado com o uso da técnica de classificação, seja possível categorizar conceitos e termos junto ao módulo de processamento de linguagem natural, que visa a compreender como tais termos estão sendo utilizados.

Destaca-se que, nessa primeira etapa, o foco não está em compreender ou em obter o significado dos termos, mas, sim, em ter uma visão do que o usuário está querendo ao fazer aquela pergunta. Nesse sentido, essa etapa tem como objetivo apenas entender a pergunta, levantando os termos principais que o usuário deseja.

Posteriormente, após essa classificação, que fornecerá os termos-chave e a identificação de qual é o tipo de pergunta que o usuário está fazendo, a próxima etapa está no **enriquecimento semântico dos termos**. Esse elemento do modelo recebe a análise prévia obtida pelo classificador, com os termos-chave levantados, para, a partir daí, obter o significado e as relações que existem entre esses termos e outros, buscando ter uma visão mais clara dos conceitos que circulam e se relacionam aos inseridos pelo usuário.

O enriquecimento semântico está vinculado a dois elementos principais, as ontologias e os vocabulários, oriundos das ferramentas da Web Semântica, e aos classificadores de texto, oriundos da camada de *machine learning*. Ontologias serão utilizadas nesse processo para possibilitar que o enriquecimento semântico leve em consideração o significado e o contexto em que os termos obtidos estão. Como relatado anteriormente, uma ontologia é capaz de expressar uma série de relações e propriedades que expressam semanticamente como um termo está posicionado frente a um determinado cenário.

O uso da ontologia é essencial nesse processo, uma vez que o desenvolvimento de ontologias é realizado considerando aspectos linguísticos e da biblioteconomia, que são capazes de expressar um domínio com clareza. No âmbito deste modelo, as principais funções das ontologias são:

- Definição de sinônimos;
- Definição de termos relacionados;
- Definição de termos genéricos e específicos;
- Definição de atributos que são relacionados a conceitos;

Adicionalmente, o enriquecimento semântico utiliza um dos principais conceitos propostos nesta tese, que está na união entre os princípios da Web Semântica com a Inteligência Artificial, por meio de técnicas de processamento de linguagem natural com processos de compreensão e classificação dos termos. Nesse contexto, o processo realizado pelo enriquecimento semântico será aprimorado com elementos classificadores de texto, que buscam, por meio de aprendizagem de máquina, com o uso de algoritmos de *topic modelling*, obter informações e classificações sobre os termos e conceitos previamente classificados pela ontologia.

Assim, a classificação ocorrerá em duas etapas, permitindo que o nível de semântica formal seja mais elevado, o que será apresentado com detalhes nas subseções seguintes, ao mesmo tempo que aumenta a probabilidade de ter pelo menos algumas informações daquele termo, seja via ontologia, seja via classificador de texto. Além disso, o uso desses classificadores é interessante, pois o uso de aprendizado de máquina tornará o processo mais aprimorado ao longo do tempo e, por outro lado, permitirá obter características das relações das ontologias.

Ao final desse processo, aquela pergunta ou termo inicial estará expandido, ampliado e enriquecido, o que é representado pelas bolinhas com o "?+" na figura 29.

A etapa seguinte está no **buscador**, que tem como função encontrar as fontes informacionais capazes de atender as necessidades informacionais dos usuários. Nesse sentido, o buscador terá três tipos principais de fontes de informação: dados publicados em formatos de *Linked Data*, fontes textuais disponíveis na Web e bases de dados estruturados diversas.

O buscador tem uma importante função, uma vez que todos os tratamentos realizados nas etapas anteriores terão como resultado, em um primeiro momento, o modo como a busca é realizada nas fontes informacionais. Assim, o buscador terá que realizar tratamento para o *Linked Data*, de como que a busca deverá ser realizada em bases RDF, utilizando linguagem SPARQL. A busca em bases de *Linked Data* é a mais completa, uma vez que possibilita utilizar as relações encontradas nas ontologias, além de permitir buscas nos elementos descritivos e expandir a sua compreensão. Isso é possível, pois OWL e RDF utilizam a mesma sintaxe, e a busca realizada em SPARQL pode utilizar todos os tipos de relações que foram identificadas na ontologia.

No que tange ao buscador, salienta-se que as fontes obtidas com o *Linked Data* serão essenciais para se ter um resultado mais satisfatório, pois as buscas realizadas nas bases do *Linked Data* terão uma grande gama de relações e propriedades que foram obtidas a partir da ontologia, e que podem ser localizadas por meio do SPARQL. Assim, os resultados obtidos

com as fontes informacionais do *Linked Data* irão auxiliar o processo de busca nas outras fontes, como os textos e as bases de dados.

Vale destacar que as buscas são realizadas utilizando tanto as informações descritivas, por meio dos metadados, quanto os próprios documentos, que permitem um melhor processo de recuperação da informação. Por tal motivo, apresenta-se no modelo a questão dos metadados, que são utilizados neste processo.

Dessa forma, serão realizadas buscas que serão aprimoradas, visando a obter resultados que sejam semanticamente aderentes àquilo que os usuários estão buscando. Isso é fundamental, pois o processo de recuperação de informação é essencial e deve ser realizado utilizando os princípios da Web Semântica e as informações extraídas das ontologias.

O módulo que aparece na sequência é o **validador**, que tem como função identificar se o processo realizado foi satisfatório. O princípio desse módulo está em utilizar as informações contextuais das ferramentas da Web Semântica e realizar uma análise das respostas com o *machine learning*, para verificar se os resultados foram satisfatórios. Caso não tenha nenhum resultado satisfatório, o processo retorna para a fase de identificação da pergunta, para que seja todo refeito. No caso de haver ao menos uma resposta, o processo segue para a etapa do filtro.

Na etapa seguinte, a do **filtro**, as respostas potenciais encontradas passam por mais um refinamento, trazendo os elementos semânticos e de classificação, que usam como base princípios de aprendizagem de máquina, para aprimorar e definir as melhores respostas para o usuário. Com a aplicação do algoritmo supervisionado de classificação, é possível obter categorias e informações mais claras das informações obtidas, auxiliando na definição daquela que pode ser mais adequada para ser a resposta dada ao usuário. Esse refinamento será apresentado nas seções seguintes. O princípio desta etapa está no uso dos significados e do contexto em que os termos se encontram para definir o que será dado como resposta para o usuário ao final do processo.

Esse processo está vinculado aos conceitos do *Question Answering*, em que o usuário irá receber uma resposta exata daquilo que está buscando. Assim, dentre os documentos e dados obtidos no buscador, deve-se realizar um processo em que é encontrado exatamente qual é a resposta daquilo que o usuário está buscando.

Esse processo é feito centrado em técnicas de processamento de linguagem natural, com o uso de palavras-chave e da proximidade entre os termos, como discutido anteriormente. No entanto, esse processo também utilizará as relações identificadas na ontologia, que serão apresentadas com detalhes nas subseções seguintes, para aprimorar os resultados alcançados, além de oferecer uma melhor compreensão do que o usuário deseja, e das informações

encontradas. As relações das ontologias fornecerão outros elementos que podem ser utilizados para a identificação de informações dentro dos documentos e das informações encontradas.

O uso das ontologias nesse processo ocorrerá para que seja possível compreender o significado dos termos, bem como as relações existentes, para que, ao identificar quais das respostas potenciais satisfazem a busca, comparar com tais relações e termos e conceitos que fornecem um nível de semântica formal e de contexto da busca realizada.

Finalmente, a última etapa do processo é o **criador da resposta**, que irá finalizar o processo fornecendo para o usuário o que foi encontrado de uma forma aderente àquilo que ele buscou. O criador de respostas irá se embasar também em bases de dados e significados para apresentar uma resposta mais aderente ao contexto do usuário, considerando os significados e sinônimos encontrados, dos termos utilizados na busca do usuário.

Essa etapa também contempla a interface de resposta que o usuário vai receber, tendo a possibilidade de retroalimentar a busca, para aprimorar, caso o usuário não tenha as suas necessidades informacionais contempladas. Esse processo interativo e iterativo aprimora os resultados alcançados, melhorando os algoritmos de aprendizagem de máquinas adotados, ao mesmo tempo que se aproxima de dar melhores resultados aos usuários.

Por fim, a última informação contida no modelo trata dos níveis de compreensão que se atinge com o modelo. Esses níveis são embasados no processamento de linguagem natural, e vai se aprofundando a partir dos módulos e do tratamento que é dado aos termos. Nesse sentido, a primeira etapa tem apenas uma compreensão morfológica e léxica, visto que se tem apenas um reconhecimento nocional de elementos da escrita ou da fala. Na sequência, os níveis sintático e semântico são utilizados para a identificação da pergunta, até a sua classificação, atingindo o enriquecimento, juntamente com o nível pragmático, em que PLN, *machine learning* e ferramentas da Web Semântica são utilizadas para atingir um nível mais aprofundado do nível de compreensão.

Outro aspecto importante a ser destacado está nos objetivos do modelo. Em suma, tem-se três objetivos principais: realizar a compreensão da pergunta e da resposta que será dada. Tal compreensão é responsável pelo entendimento da pergunta, além de trazer a necessidade de uma interface que irá criar a resposta que atende as necessidades informacionais dos usuários. Esse objetivo tem como foco principal o processamento de linguagem natural para permitir que a pergunta seja compreendida de forma eficiente.

O segundo objetivo está vinculado à semântica formal. O foco está em obter qual é o significado daquelas informações trazidas a partir do tratamento da pergunta, relacionando os conceitos que as técnicas de Inteligência Artificial estão encontrando, juntamente com a

semântica relacionada às ferramentas da Web Semântica. Esse mesmo ponto é abordado no que se refere à definição do que será dado como resposta para os usuários, realizando os relacionamentos das informações que lhes serão apresentadas.

Finalmente, o terceiro objetivo é a recuperação da informação, em que serão levantadas as informações que atendem às necessidades informacionais dos usuários, levando em consideração os aspectos da análise semântica realizada, tanto no aspecto da IA, quanto das ferramentas da Web Semântica. Devido a tal objetivo, realiza-se a busca das informações e de tratamento das respostas obtidas, visando a passar para as fases seguintes do processo.

A seguir apresentam-se, com detalhes, todas as camadas e módulos que fazem parte do modelo proposto. Essas camadas serão exploradas com detalhes, com sub arquiteturas, com detalhes de como serão feitas e o que embasa teórica e praticamente cada uma delas. Assim, todo o processo será explorado e apresentado na sequência.

### 5.3.1 Camadas de suporte

O modelo proposto possui, além dos diversos módulos, três camadas que auxiliam todos os seus processos As três camadas são: i) Processamento de linguagem natural, responsável por tratar e realizar a compreensão dos textos em linguagem natural; ii) *Machine learning*, responsável por obter informações dos textos e aprender junto com o processo do modelo e; iii) Ferramentas da Web Semântica, responsável por auxiliar na definição dos significados e dos contextos dos termos e respostas que se vinculam ao modelo proposto.

A seguir apresentam-se detalhes de cada camada, no âmbito do modelo proposto.

#### 5.3.1.1 Processamento de linguagem natural

O processamento de linguagem natural é parte central do modelo para a compreensão da linguagem natural nos processos computacionais. Essa compreensão é fundamental em diversas partes do modelo, desde a compreensão da pergunta, passando pelo aprimoramento dos resultados obtidos pelo buscador, chegando até a construção da resposta.

Os diversos momentos em que o processamento de linguagem natural é utilizado são detalhados na seção de funcionamento do modelo. Em suma, essa camada auxilia na identificação dos conceitos em textos em linguagem natural, realiza a extração de palavras-chave, faz a análise de textos obtidos em possíveis resultados, trata da extração dos principais

conceitos de um texto para a construção da resposta e permite o entendimento de conceitos e do seu significado nos diversos textos.

Para a realização de todos esses processos, a camada de processamento de linguagem natural desempenha algumas funções que serão utilizadas por vários dos módulos do modelo. O quadro 6 apresenta as principais funções desempenhadas pela camada de PLN, o seu funcionamento e alguns exemplos de ferramentas que podem ser utilizadas para a realização dessas funções.

**Quadro 6** – Funções realizadas pela camada de processamento de linguagem natural

| **Função** | **Descrição** | **Exemplo** |
|---|---|---|
| Identificação de conceitos e entidades e classificação | Esse processo acontece com a identificação de termos que se referem a conceitos em um texto. A ideia está em extrair de um texto os principais conceitos que estão vinculados a ele, visando obter do texto os termos mais importantes. Esse processo acontece por meio de uma subárea de processamento de linguagem natural chamada de reconhecimento de entidades nomeadas, que identifica, por meio de regras ou de aprendizagem de máquinas, quais são os conceitos e as entidades presentes naquele texto. Neste trabalho, esse processo é fundamental para identificar quais são as entidades e os conceitos presentes em uma pergunta ou em um texto. A partir do texto, são encontrados e anotados esses termos, que são utilizados posteriormente.<br>Exemplo: Da frase "O livro Dom Casmurro escrito por Machado de Assis foi publicado em 1899", é possível extrair conceitos e classificá-los. No caso, "Dom Casmurro" seria identificado como um conceito que se refere a um livro, "Machado de Assis" seria identificado como uma entidade que se refere a um autor e o "1899" seria identificado como um entidade que se refere a um ano. | Há diversas bibliotecas e ferramentas que realizam esse tipo de processamento, como o IBM Watson, com a sua função de *Natural Language Understanding*, a Amazon, com o Amazon *Comprehend*, o Google Cloud, com o serviço *Natural Language*, por bibliotecas em Python como o NLTK e pela construção de algoritmos que realizam esse procedimento, seja com o uso de regras ou de *Machine Learning.* |
| Realização de análise semântica dos termos | A análise semântica dos termos é essencial neste projeto, uma vez que alguns aspectos do entendimento da semântica formal dos textos parte da linguagem natural, para posteriormente ser inserido no contexto das ferramentas da Web Semântica. A realização dessa análise é uma consequência da identificação de conceitos e entidades, aliados a análises morfológicas e sintáticas realizadas, que permitirão a compreensão das relações entre os conceitos.<br>Esse processo acontece após definir e etiquetar as entidades, e buscará, com o apoio dos verbos, definir como esses conceitos estão se relacionando, visando a encontrar o seu sentido. Junto a isso, a classificação das entidades, como relatado | As mesmas bibliotecas e os serviços apontados para a identificação de entidades podem ser utilizados nesse processo. Isso porque, a realização dessa análise irá utilizar a identificação dos termos, a análise sintática e a interpretação das categorias a qual elas pertencem. |

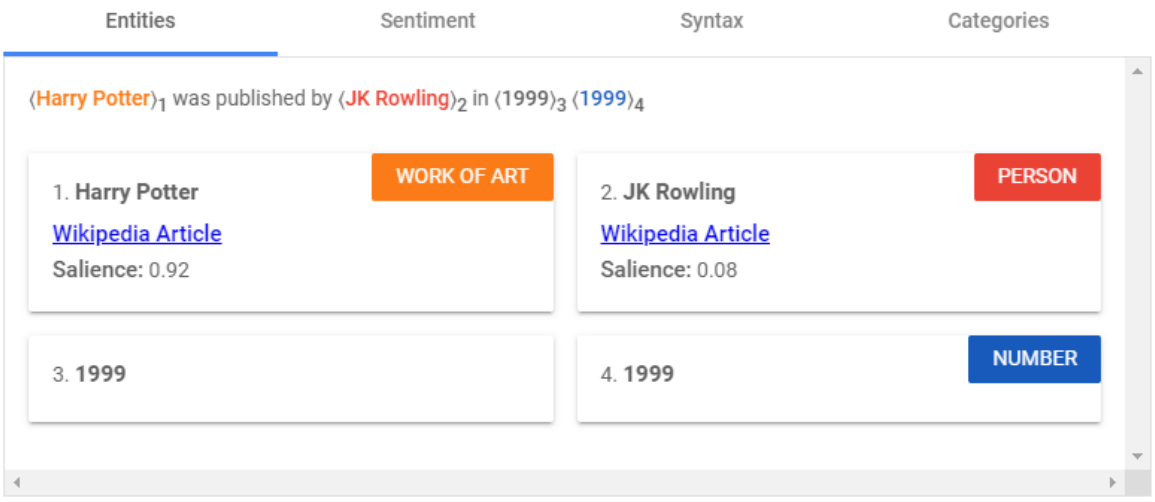
| | | |
|---|---|---|
| | anteriormente, classificando um conceito como um determinado tipo é essencial. Por exemplo, a classificação de Dom Casmurro como um livro. Assim, no exemplo anterior "O livro Dom Casmurro escrito por Machado de Assis foi publicado em 1899", teria além das classificações de entidades, a relação de que "Dom Casmurro" foi escrito por "Machado de Assis" e que "Dom Casmurro" foi publicado em "1899", tirando assim, duas relações com o potencial semântico para a sua interpretação. | |

Fonte: elaborado pelo autor.

O quadro 6 revela diversos aspectos tratados pela camada de processamento de linguagem natural que compõem o modelo proposto.

Um dos aspectos evidenciados pelo quadro 6 está na identificação de entidades e conceitos. Um exemplo desse processo pode ser observado na figura 30, que explicita como um serviço desse tipo se comporta. No caso desta figura, utilizou-se o serviço do Google Cloud como demonstração.

**Figura 30** – Demonstração da identificação de conceitos a partir de um texto




Fonte: Resultado de consulta realizada no Google Cloud em 24 fev. 2020.

Na análise realizada é possível ver como as entidades são etiquetadas a partir do texto "*Harry Potter was published by J. K. Rowling in 1999*", destacando diversos elementos no texto.

Já no que se refere à análise semântica, a partir dessa classificação realizada, compreendendo que Harry Potter é um livro, JK Rowling é uma autora e 1999 é um ano, utiliza-

se a análise sintática, pela qual se identifica a ligação entre as três entidades para conseguir encontrar esse relacionamento entre os diversos conceitos.

Uma camada que está vinculada a de processamento de linguagem natural é a de *machine learning*, que será apresentada a seguir.

#### 5.3.1.2 *Machine Learning*

A camada de *machine learning* é responsável por auxiliar nos processos de processamento de linguagem natural e Web Semântica, tornando a identificação de conceitos e elementos mais precisa, bem como aprimorando o processo de classificação dos relacionamentos entre os termos e a semântica das frases. Essa camada será responsável por permitir que o modelo aprimore com o tempo, de acordo com as respostas dadas e o atendimento ou não às necessidades informacionais dos usuários.

A seguir, no quadro 7, apresentam-se alguns algoritmos e funções que essa camada utilizará nos diversos módulos que compõem o modelo.

**Quadro 7** – Funções da camada de *machine learning*

| **Função/Algoritmo** | **Descrição** | **Exemplo** |
|---|---|---|
| *Topic modelling* | O *topic modelling* é um método de análise de grandes quantidades de textos, que visa sumarizar e encontrar os termos-chave de uma coleção de documentos, sem ter necessariamente conhecimento prévio deles. Esse método utiliza um conceito semântico, de buscar entender a relação existente entre os termos, embasando-se na ocorrência frequente de termos próximos.<br>No caso do presente trabalho, a abordagem de *topic modelling* pode ser utilizada para obter informações de fontes informacionais ou de textos vinculados a um determinado domínio, para que seja utilizada como fonte de informação para a definição de relacionamentos e na construção de respostas para ser oferecida aos usuários.<br>Um exemplo seria realizar as buscas para a determinação das respostas dadas aos usuários em uma base de artigos científicos; a análise de *topic modelling* poderia auxiliar na compreensão daquele domínio, ao mesmo tempo que facilitaria uma posterior recuperação da informação no domínio determinado. | Bibliotecas em Python utilizando *Latent Dirichlet Allocation* (LDA) pode ser utilizado para esse processo, oferecendo uma análise satisfatória. |
| *Dependency structure* | O *dependency structure* é um meio de realizar a análise de textos por meio da gramática existente, utilizando as regras definidas. Esse método pode ser utilizado com o apoio de *machine learning*, que favorece a realização de desambiguação e aprimora | O uso de algoritmos como *Principal Component Analysis* (PCA) pode ser utilizado com o fim |

| Função/Algoritmo | Descrição | Exemplo |
|---|---|---|
| | as análises do texto, visando a encontrar o relacionamento entre os diversos termos de um texto. Destaca-se ainda que o uso de *dependency structure* auxilia na interpretação semântica de gramáticas, aprimorando o encontro das relações existentes entre os termos.<br>No âmbito deste trabalho, a interpretação das perguntas realizadas pelos usuários pode ser significativamente aprimorada com esses algoritmos, uma vez que permitirá o uso de gramáticas bem expressivas para entender o que o usuário quer, a partir das relações existentes entre os diversos termos da pergunta. | de realizar os processos de *dependency structure*. |
| Apoio à classificação das entidades | Outro importante aspecto trabalhado na camada de *machine learning* é o apoio à classificação das entidades, vinculadas à camada de processamento de linguagem natural. Tal apoio utiliza os algoritmos supracitados, junto com as bases de dados que serão utilizadas para a classificação de entidades. Em suma, nesse processo ocorre o treinamento de determinadas entidades com sua respectiva classificação, mas tem-se também um contínuo aprimoramento, vinculado ao uso que se tem do modelo. | O uso dos algoritmos citados, junto com bases da IBM Watson e Google Cloud auxiliam neste processo, uma vez que existem elementos que facilitam essa classificação dos termos. |

Fonte: elaborado pelo autor.

O quadro 7 apresenta alguns dos algoritmos e tipos de uso do *machine learning* no âmbito do modelo. Há uma relação clara entre a adoção dos algoritmos de *machine learning* e processamento de linguagem natural, uma vez que o modelo se embasa a todo tempo nas análises de conteúdos textuais.

O uso desta camada torna o modelo mais inteligente, visto que favorecerá uma evolução constante dos módulos e das análises, permitindo que as análises sejam aprimoradas ao longo do tempo. Junto a essa camada e à camada de processamento de linguagem natural, tem-se as ferramentas da Web Semântica, que serão exploradas na sequência.

#### 5.3.1.3 Ferramentas da Web Semântica

As ferramentas da Web Semântica serão utilizadas na maioria dos módulos, para possibilitar que as análises alcancem um nível mais elevado de semântica formal, possibilitando a compreensão dos significados e do contexto dos termos e das suas relações. Os instrumentos da Web Semântica permitem a definição dos domínios que serão utilizados, além de permitir

buscas em bases de dados estruturadas, como o *Linked Data*, permitindo a obtenção de resultados mais expressivos.

O quadro 8 apresenta as ferramentas que serão utilizadas, sua descrição e um exemplo de uso.

**Quadro 8** – Camada de ferramentas da Web Semântica

| **Ferramenta/Função** | **Descrição** | **Exemplo** |
|---|---|---|
| Ontologias, tesauros e vocabulários | As ontologias buscam auxiliar na compreensão por parte dos mecanismos computacionais do significado e do contexto em que os termos estão inseridos. Dessa forma, as ontologias são utilizadas como um instrumento que auxiliam na descrição dos domínios, permitindo, quando aliados às tecnologias de processamento de linguagem natural, aproximar a linguagem computacional da linguagem natural.<br>No presente trabalho, as ontologias serão utilizadas como instrumentos que permitem a compreensão dos domínios que serão trabalhados no contexto da busca. Além disso, as ontologias serão utilizadas em diversos momentos do modelo, como a expansão da busca, uma vez que, a partir dos termos que o usuário escolhe na pergunta, ocorre um enriquecimento desses termos com relação a outros termos, e a validação, se os resultados da busca forem aderentes ao cenário que o usuário está buscando; para isso, as informações recuperadas são comparadas com o contexto da ontologia. | Podem-se utilizar ontologias em OWL, encontradas em diversos buscadores de ontologias. Adicionalmente, destaca-se que há uma grande quantidade de ontologias OWL já construídas em diversos domínios. |
| Motor de busca SPARQL | O uso de ontologias e de fontes informacionais baseados no *Linked Data* torna necessário que haja mecanismos para recuperação da informação nesses instrumentos baseados nos princípios da Web Semântica. Assim, o uso do SPARQL é fundamental para que o encontro das relações e a expansão da busca aconteça de forma mais eficiente.<br>O SPARQL, como uma linguagem baseada no RDF, permite que a expressão e a semântica dos instrumentos sejam potencializadas, uma vez que não ocorre uma diferença no paradigma em que as informações são armazenadas e na maneira como elas são buscadas. | O Apache Jena pode ser utilizado para a construção do motor do SPARQL, permitindo a criação das consultas de forma dinâmica. |

| Ferramenta/Função | Descrição | Exemplo |
|---|---|---|
| | No âmbito do trabalho, o uso de um motor de busca utilizando o SPARQL é essencial, visto que há regras a serem consideradas nos diferentes momentos em que as ontologias e o *Linked Data* são utilizados, exigindo que as buscas nesses instrumentos aconteça de acordo com as regras do modelo. | |
| *Linked Data* | As fontes informacionais utilizando os princípios do *Linked Data* apresentam uma alta expressividade e podem, quando aliadas com outras fontes, permitir que o processo de recuperação da informação seja mais eficiente. Isso porque, com a recuperação de informações do *Linked Data*, sejam informações descritivas ou os próprios documentos, tem-se uma base estruturada que poderá auxiliar a comparar com os resultados obtidos em outras fontes, e mesmo utilizar o que se recuperou do *Linked Data* para realizar uma recuperação mais aprimorada, que utilize os elementos encontrados para enriquecer a busca. | O *Linking Open Data* reúne diversas bases de dados seguindo os princípios do *Linked Data*, podendo ser utilizadas como fontes para o modelo. |

Fonte: elaborado pelo autor.

O quadro 8 apresenta as principais contribuições das ferramentas da Web Semântica no modelo. Essas ferramentas estão vinculadas aos diversos módulos, sendo fundamentais para se atingir os seus objetivos.

Na sequência, apresenta-se o funcionamento do modelo proposto.

### 5.3.2 Funcionamento do modelo

O modelo foi dividido em sete módulos principais: identificador da pergunta, classificador da pergunta, enriquecimento semântico da pergunta, buscador, validador, filtro e criador da resposta. Cada módulo será explorado com detalhes a seguir.

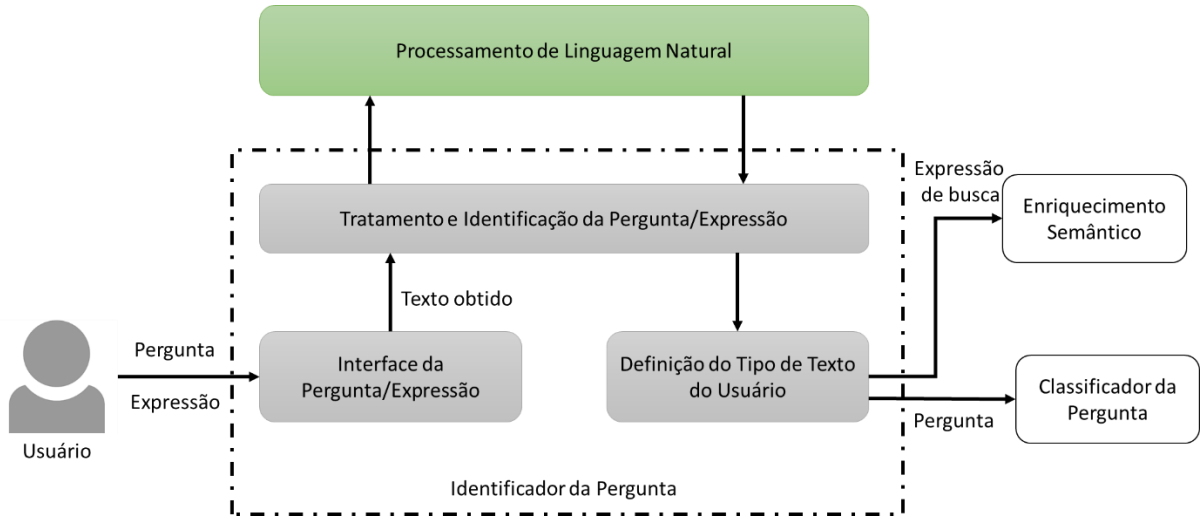
#### 5.3.2.1 Identificador da pergunta

O primeiro módulo do modelo apresentado é o identificador da pergunta, que tem como objetivo verificar se aquilo que um usuário inseriu no sistema é uma pergunta ou palavras-chave de uma busca. Esse módulo é fundamental devido a necessidade de torná-lo capaz de tratar tanto perguntas que necessitam ser entendidas, quanto palavras-chave, caso o usuário não deseje utilizar uma interface que seja orientada a perguntas.

Dessa forma, basicamente, o identificador irá verificar se aquilo que foi inserido por um usuário tem um formato de pergunta ou não. Para isso, utiliza-se a camada de processamento de linguagem natural para apoiar esse processo.

Os detalhes do funcionando desse módulo são explicados na figura 31, em que se apresenta a sua arquitetura.

**Figura 31** – Módulo identificador da pergunta




Fonte: elaborado pelo autor.

A figura 31 demonstra que a partir da inserção do texto pelo usuário, que pode ser uma pergunta ou uma expressão de busca, ocorre um processo de tratamento e de identificação daquele texto. Dessa forma, esse módulo utiliza processamento de linguagem natural para identificar os elementos que podem caracterizar aquele texto como uma pergunta, tais como o uso de pronomes relativos (como, o quê, qual, entre outros), o uso do ponto de interrogação e a estrutura da escrita/fala.

A partir disso, há uma definição do tipo de texto do usuário e, caso o módulo tenha identificado que o texto é uma pergunta, o modelo segue para o Classificador da Pergunta. No entanto, caso o modelo tenha identificado que era uma expressão de busca, o modelo segue diretamente para o Enriquecimento Semântico, pois, como não é uma pergunta, não há a necessidade de classificá-la.

A seguir apresenta-se o módulo do classificador da pergunta.

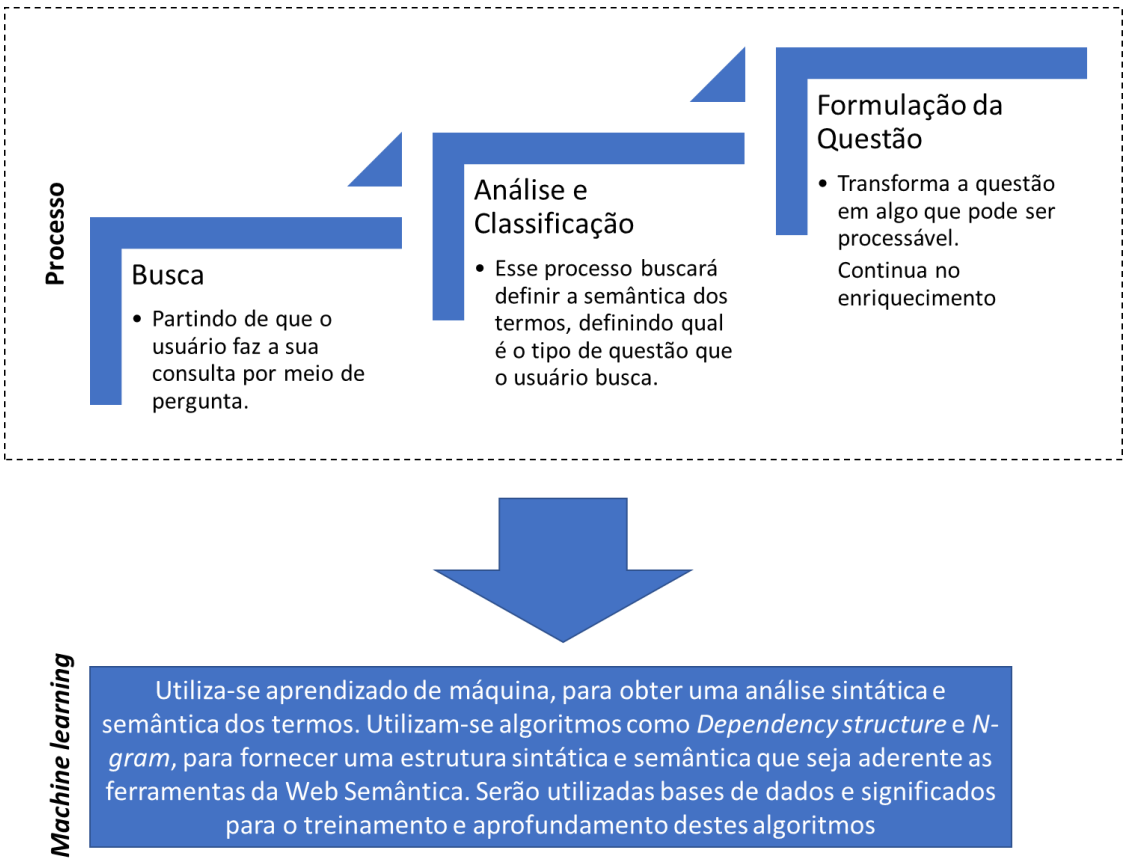
### 5.3.2.2 Classificador da pergunta

O classificador de pergunta tem um papel central no modelo, no que tange às questões do processamento de linguagem natural para que, primeiramente, seja capaz de compreender a pergunta realizada pelo usuário, além de conseguir estruturar a resposta que atenda ao que o usuário está desejando.

Esse classificador de perguntas tem como responsabilidade a compreensão tanto do que o usuário está buscando, quanto do impacto e da forma como a resposta será dada para o usuário. Nesse contexto, utilizam-se nesse módulo os procedimentos da área de processamento de linguagem natural para, entrando no nível semântico do PLN, extrair informações importantes para a compreensão do tipo de pergunta feita pelo usuário, o que influencia na resposta dada a ele ao final do processo. Outro vínculo está com a camada de *Machine Learning* que, por meio da pergunta, deverá utilizar os seus procedimentos para compreender e classificar os termos que compõem a pergunta, utilizando algoritmo supervisionado de classificação.

Na figura 32, explica-se como é o processo de classificação da pergunta que acontece nesse módulo.

**Figura 32** - Processo do classificador da pergunta



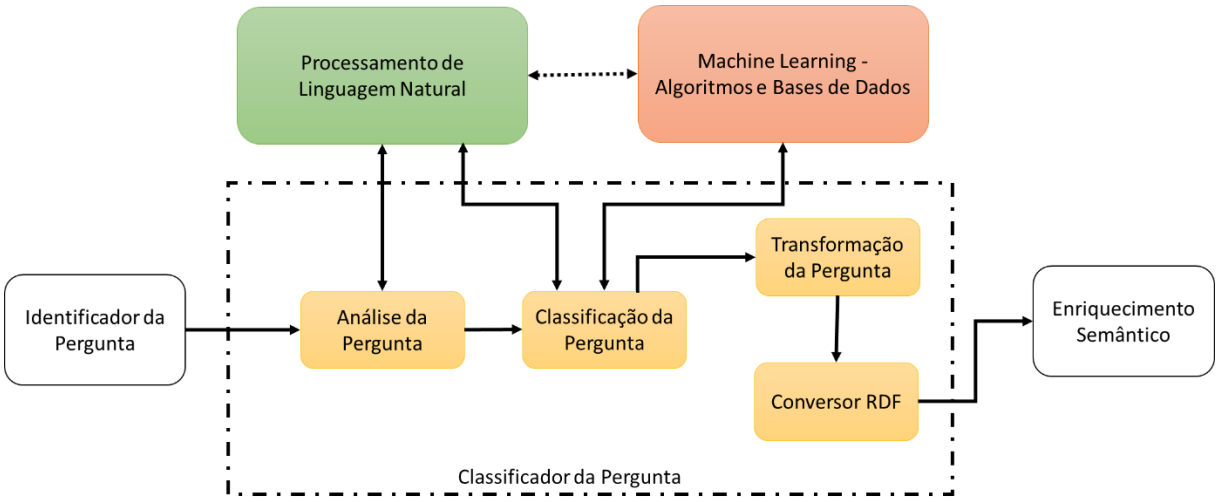

Fonte: elaborado pelo autor.

Em suma, essa figura demonstra qual é o passo a passo para a realização do processo da pergunta realizada pelo usuário, cujo enfoque é a utilização de algoritmos de *machine learning*, junto ao processamento de linguagem natural, que permitirão alcançar o nível de sintática e semântica da análise realizada. Destaca-se ainda que a saída desse módulo será realizada permitindo que os resultados obtidos com o processamento da pergunta sejam estruturados em RDF, linguagem base para processar as informações da Web Semântica.

Essa estruturação em RDF permitirá que os dados possam ser mais adequados às camadas posteriores, em que haverá a necessidade de comparar e de utilizar as ontologias, para aprimorar os resultados alcançados.

A seguir, na figura 33, apresenta-se a arquitetura do item compreensão da pergunta, que foi apresentado no modelo geral.

**Figura 33** – Arquitetura do módulo classificador da pergunta




Fonte: elaborado pelo autor.

A figura 33 explicita o funcionamento do módulo classificador da pergunta, relacionado com as camadas de processamento de linguagem natural e *machine learning*. Após a identificação da pergunta, o primeiro processo está na realização da sua análise, que visa a obter algumas informações importantes desse elemento, especialmente o foco que a pergunta possui. Esse primeiro elemento está vinculado ao processamento de linguagem natural, para conseguir tratar e analisar as características do texto.

Na sequência, tem-se o classificador da pergunta, responsável por identificar qual é o tipo de pergunta, além de ser responsável por identificar qual será o tipo da resposta que deverá

ser fornecido. Esse elemento utiliza, além do processamento de linguagem natural, os algoritmos de *machine learning*, que irão auxiliar na identificação desses elementos, analisando outras respostas, dados históricos e textos que podem auxiliar a tornar mais aderente a resposta fornecida.

Após essa etapa, a transformação da pergunta buscará encontrar os termos mais relevantes daquela pergunta, chamados de foco da pergunta, que serão utilizados para serem enriquecidos na sequência. Esse elemento é uma continuidade da classificação da pergunta, mas com um enfoque maior na definição dos termos que serão utilizados para o processo de recuperação da informação em si.

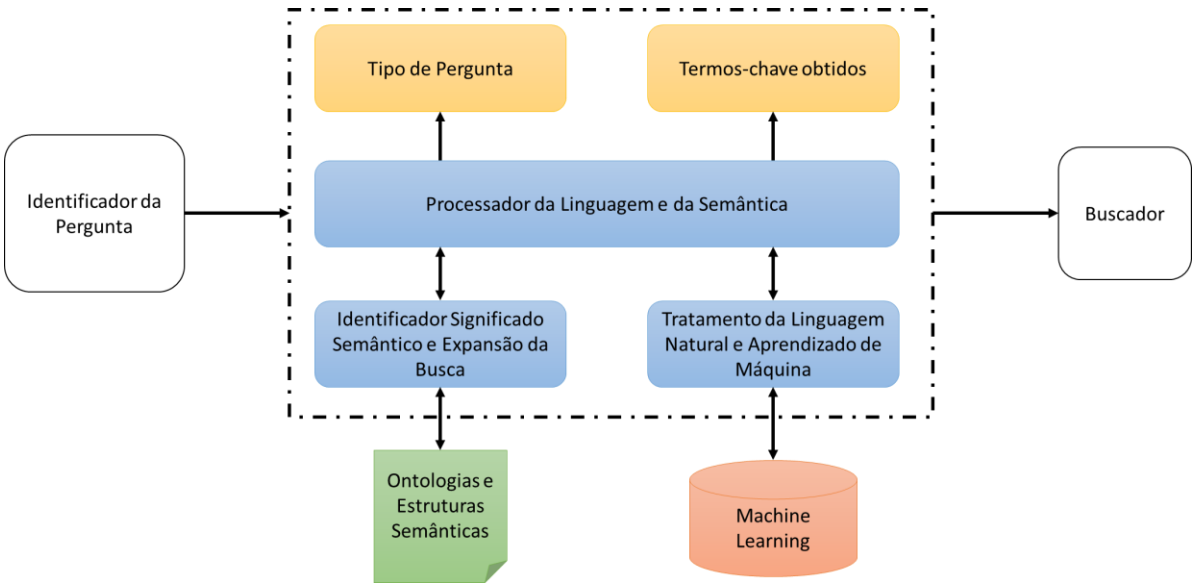
Por fim, no conversor RDF, realiza-se o processo de conversão dos resultados alcançados com os algoritmos em dados estruturados, seguindo o formato do RDF. A partir disso, segue-se para o enriquecimento semântico.

#### 5.3.2.3 Enriquecimento semântico dos termos

O módulo de enriquecimento semântico dos termos é fundamental para o modelo, uma vez que é nesse módulo que se realiza o processamento dos termos e a identificação da semântica, usando tanto as ferramentas da Web Semântica quanto as ferramentas do PLN e da Inteligência Artificial.

Esse módulo recebe do classificador da pergunta ou identificador da pergunta uma estrutura em formado RDF, que deverá ser tratada de modo a estruturar as formas para a realização da busca, seguindo as ferramentas da Web Semântica. Apresenta-se abaixo, na figura 34, a arquitetura do enriquecedor semântico dos termos.

**Figura 34** – Arquitetura do enriquecedor semântico dos termos




Fonte: elaborado pelo autor.

A arquitetura apresentada parte do tratamento e dos significados encontrados na etapa de compreensão para, ao final, transformar isso em termos de busca que serão utilizados para realizar a recuperação da informação no buscador. Destaca-se que, no processo de enriquecimento semântico dos termos, tanto as informações do tipo da pergunta, quanto dos termos-chave obtidos, serão utilizadas para a realização do processo.

No que tange aos aspectos de *machine learning*, utiliza-se classificação de termos e conceitos, que irão aprendendo de acordo com a utilização; além disso, usam-se bases de dados que possuem um conjunto grande de termos e significados. Assim, algoritmos de aprendizagem supervisionado de classificação e algoritmos não supervisionados como o *topic modelling* são capazes de aprimorar esse enriquecimento, ao encontrar relações entre os termos e promover o seu entendimento.

Ressalta-se que o processo parte das informações das perguntas, para buscar uma compreensão do significado, tendo como base as ontologias, e realizando um tratamento de linguagem natural, que utiliza aprendizagem de máquina. Dessa forma, o enriquecedor expande o significado que foi obtido a partir da pergunta.

Vale destacar que um importante aspecto desse módulo é o processo de expansão de busca, em que, a partir do que foi extraído da pergunta, utilizam-se as ontologias para aumentar o nível de semântica, obtendo informações que estão vinculadas a esse processo. Para isso, tem-se como base o estudo feito por Coneglian (2017), em que foi caracterizada a maneira como as ontologias podem ser utilizadas para a recuperação da informação.

O quadro 9 demonstra algumas das propriedades de classes das ontologias que são utilizadas para o processo de recuperação da informação. As divisões apresentadas são relacionadas às propriedades das classes que impactam no modo como os termos se relacionam. Além disso, apresenta-se a ação que vincula os diversos termos, além da ordem em que os termos são vinculados. Destaca-se que a ordem é importante, devido a possibilidade de haver diversos termos obtidos a partir das relações das ontologias, sendo necessário a definição da ordem em que tais termos serão relacionados.

**Quadro 9** - Ações das propriedades de classes

| **Divisão** | **Ação** | **Ordem** |
|---|---|---|
| Aproximação | ((X) AND Y) | 2 |
| Hierarquia (ou especificação) | ((X) OR Y) | 1 |
| Relacionamento entre diversos conceitos relacionados | ((X) AND (Y1 OR Y2 OR Y3… OR Yn)) | 3 |
| Igualdade | ((X) OR Y) | 1 |
| Intersecção | (X) AND Y | 2 |
| União | (X) OR Y | 1 |
| Diferenciação | ((X) NOT AND (Y)) | 4 |

Fonte: Coneglian (2017, p. 87).

O quadro 9 aponta como os termos encontrados nas ontologias serão tratados no âmbito da camada de compreensão, sendo importante para definir a nova expressão de busca construída, após a expansão dos termos.

A partir do processo de enriquecimento dos termos, o buscador, que será apresentado a seguir, poderá encontrar os documentos e informações que visam a atender, de forma mais eficiente, as necessidades informacionais dos usuários.

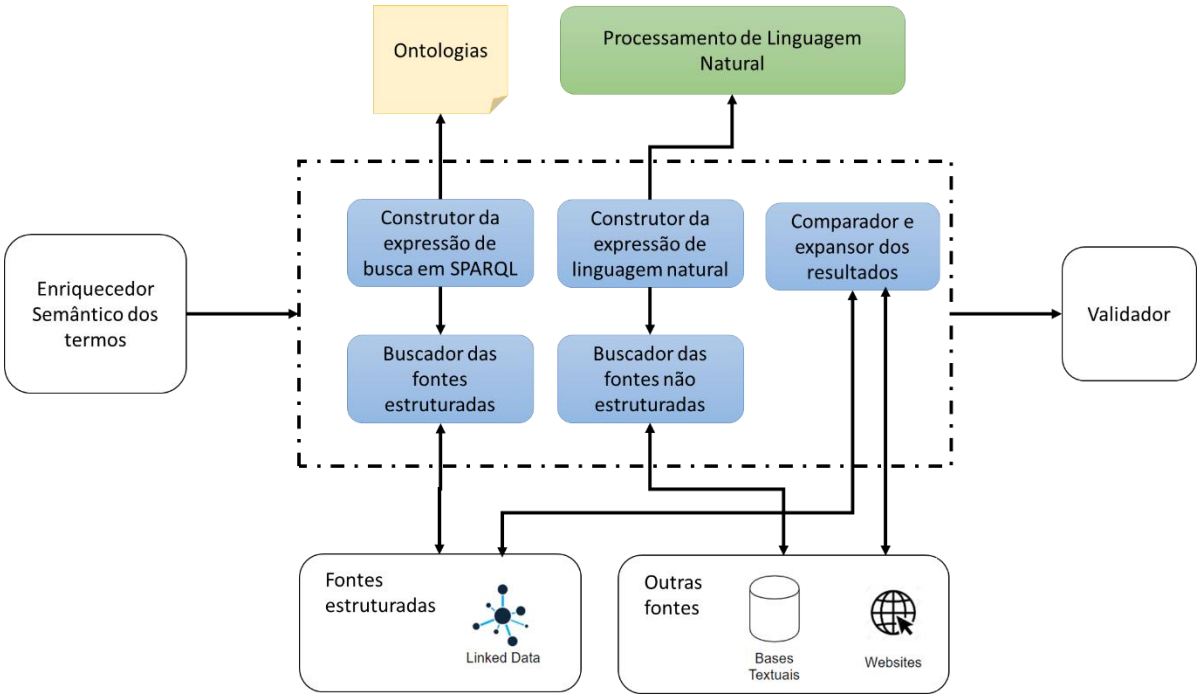
#### 5.3.2.4 Buscador

O buscador é responsável pela recuperação, no modelo proposto, de dados de diversas fontes informacionais, inclusive dados que seguem os formatos do *Linked Data*. O uso do *Linked Data* como uma das fontes informacionais foi adotado, pois a estrutura semântica de outros módulos é potencializada quando se utilizam dados que seguem um formato semântico mais aderente às ferramentas da Web Semântica.

Para tanto, utiliza-se o modelo proposto por Coneglian (2017), em que todas as propriedades das ontologias são associadas a ações dentro da recuperação da informação. A expansão da busca já aconteceu no contexto do enriquecedor semântico, utilizando o modelo supracitado, porém, nessa camada as possibilidades de uso da ontologia podem ser importantes para que haja um melhor relacionamento entre os dados coletados a partir das fontes que seguem os princípios do *Linked Data* e o contexto obtido a partir das ontologias.

O elemento desta camada é o buscador, que, partindo dos termos expandidos da camada anterior, realiza o processo de busca. No entanto, essa busca utiliza as ferramentas da Web Semântica, além de outros processos para que continue a apresentar um nível mais elevado de semântica formal. A figura 35 apresenta a arquitetura do módulo buscador.

**Figura 35** – Arquitetura do módulo buscador




Fonte: elaborado pelo autor.

A arquitetura da figura 35 foi pensada de modo que os termos expandidos estejam na entrada do módulo, saindo, ao final, os resultados em RDF. O módulo apresenta dois tipos principais de fontes informacionais, as fontes estruturadas e outras fontes, que irão impactar a maneira como a busca é realizada. Primeiramente, quanto as fontes estruturadas, o módulo funciona de modo que, partindo dos termos expandidos, seja criada a expressão de busca utilizando o SPARQL. Para a construção dessa expressão de busca, esse construtor utilizará ontologias, que poderão auxiliar na definição de como as relações podem ser mais bem exploradas na consulta. Além disso, a conversão dos termos expandidos usa a estrutura das ferramentas da Web Semântica, fazendo com que essa conversão para o SPARQL não perca a expressividade semântica.

Após a criação da expressão, é realizada a busca em si, em que a expressão do SPARQL consulta as diversas bases de dados estruturados, obtendo, assim, resultados que atendam as expressões. A ideia desse processo está em obter diversos resultados, pois haverá outros módulos após esse que poderão filtrar a melhor resposta para o usuário. Em suma, nesse processo devem ser encontradas várias respostas que podem atender as necessidades informacionais dos usuários, para que depois as etapas seguintes possam trabalhar com mais eficiência no melhor resultado.

No que se refere as fontes não estruturadas, utilizam-se os termos enriquecidos para a construção das expressões de busca. Para isso, tem-se o apoio da camada de processamento de linguagem natural, que irá fornecer recursos para a eliminação de termos sem expressividade, entre outros aspectos. A partir desse processo, com os termos que serão buscados, utiliza-se o buscador de fontes não estruturadas, que, em suma, irá buscar, em websites e bases de dados, respostas adequadas para a pergunta realizada pelo usuário.

Após a obtenção dos resultados, tanto das fontes estruturadas, quanto das não estruturadas, realiza-se uma comparação com as ontologias e com o processo da expansão da expressão, de modo que se possa verificar se alguns dos resultados não foram obtidos das bases de dados. Dessa forma, busca-se retornar às fontes informacionais visando a encontrar novas informações que possam atender ao que os usuários esperam, ocorrendo alguns ciclos de consultas às fontes de *Linked Data* e a outras fontes informacionais.

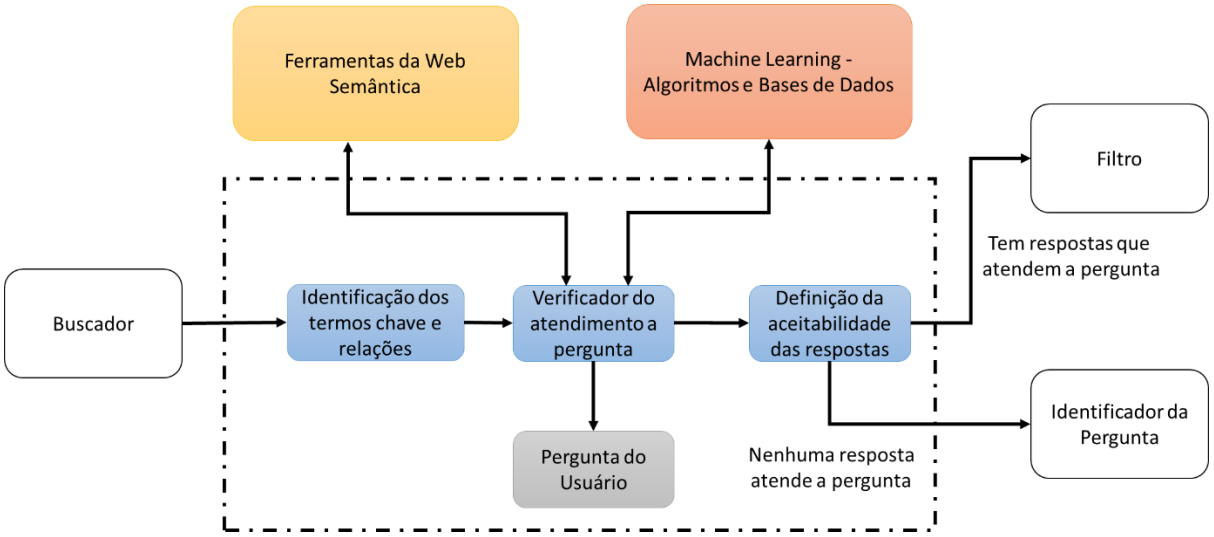
Por fim, tem-se os resultados estruturados em RDF e tais resultados devem conter a resposta que melhor atenda a pergunta do usuário.

#### 5.3.2.4 Validador

O validador é o módulo responsável por analisar e verificar se as respostas obtidas são aceitáveis para formular, posteriormente, a resposta dada ao usuário. Esse módulo irá definir se o processo deverá ser todo refeito, para utilizar outros elementos, outras relações das ontologias e outros termos de busca, caso o resultado não seja satisfatório, ou se poderá seguir com aquelas possíveis respostas.

O módulo está vinculado à camada de *machine learning* e de ferramentas da Web Semântica, visando a analisar todos os resultados de acordo com os algoritmos e com a compreensão do contexto que é possível obter com aqueles elementos. A figura 36 apresenta a arquitetura do módulo validador.

**Figura 36** – Arquitetura do módulo validador




Fonte: elaborado pelo autor.

A arquitetura desse módulo apresentada na figura 36 detalha o seu funcionamento. A partir das respostas estruturadas, há o processo de identificação dos termos-chave e das relações existentes de cada resposta, que servirá para o modelo entender cada resposta, para que possa avaliar se é aderente ou não à pergunta. Assim, na sequência, há o processo de verificação se as respostas atendem ou não à pergunta do usuário, utilizando tanto as ferramentas da Web Semântica, quanto os algoritmos e bases de dados do *machine learning* e tendo acesso, também, à pergunta inicial feita pelo usuário.

O uso dessas duas camadas nesse processo ocorre devido à necessidade de se ter elementos para utilizar como base para comparar se os resultados obtidos estão aderentes à

pergunta inicial. Essas camadas, com o auxílio das ontologias e de comparações realizadas utilizando os algoritmos de *machine learning*, permitem a comparação e a análise.

O *machine learning* utilizará algoritmos de classificação que permitam determinar se os resultados obtidos estão inseridos em grupos que atendem ou não aos resultados esperados para ser composta a resposta dada ao usuário. Nesse sentido, esses algoritmos e os resultados obtidos poderão demonstrar a aderência que os resultados oriundos do buscador possuem, tornando o processo mais inteligente e semântico, com o apoio das ferramentas da Web Semântica.

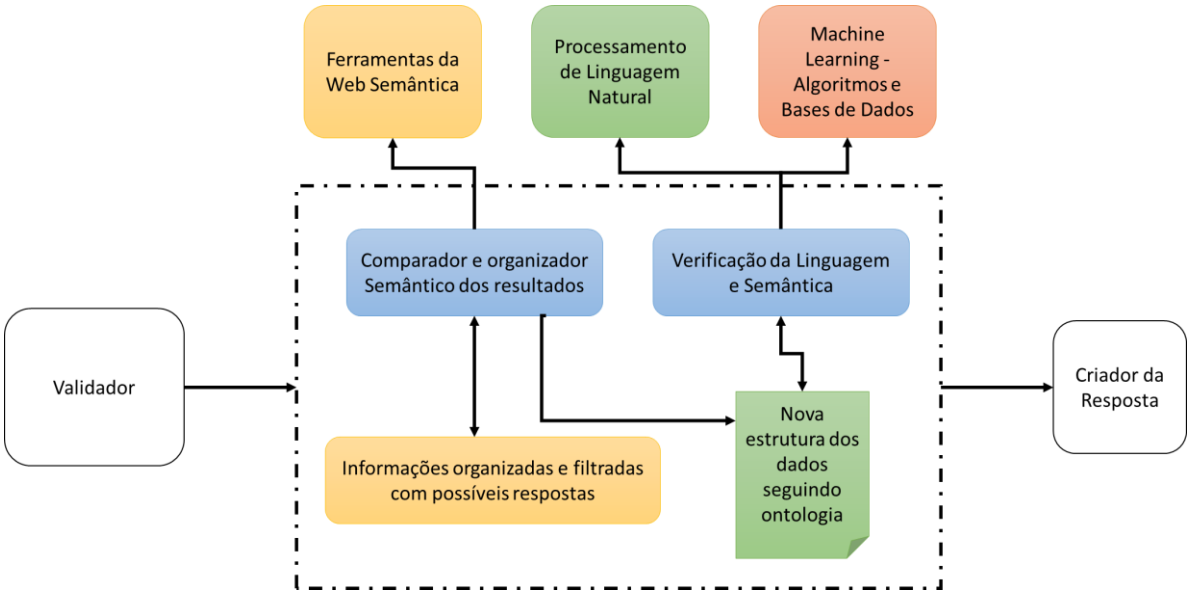
Por fim, esse último elemento, o verificador, irá reconhecer se cada resposta está ou não aderente à pergunta, possibilitando definir, de acordo com a aceitabilidade da resposta, se a próxima etapa é o filtro, caso tenha várias respostas que atendem a pergunta, ou se é o identificador da pergunta, voltando ao início do processo do modelo, caso não haja respostas aderentes.

A seguir apresenta-se o módulo do filtro, que é a etapa seguinte ao verificador.

#### 5.3.2.5 Filtro

O módulo de filtro é referente às respostas que já foram obtidas e processadas e devem ser classificadas e filtradas. Dessa forma, esse módulo irá, a partir dos resultados obtidos das bases de *Linked Data* e de outras fontes informacionais, processar as informações, para que se tenha resultados aderentes, com um nível de semântica mais elevado, e que sigam em formato RDF, como demonstrado na figura 37.

**Figura 37** – Arquitetura do módulo de filtro




Fonte: elaborado pelo autor.

A arquitetura apresentada na figura 37 apresenta dois elementos principais, o comparador e organizador semântico dos resultados e a verificação da linguagem e semântica. Ambos buscam tratar os resultados de modo que, ao final, obtenham-se resultados com mais expressividade e que possam atender com mais eficiência o usuário.

O processo está organizado de modo que as informações vindas do validador, que contém as respostas, passarão por um processo no comparador e organizador semântico dos resultados. Em suma, esse módulo irá utilizar as estruturas das ontologias para filtrar e organizar os resultados, seguindo a estrutura da ontologia, buscando encontrar informações que foram encontradas, mas que não estão aderentes ao contexto obtido pela ontologia. Esse módulo servirá como uma filtragem semântica, em que os resultados obtidos, que não seguirem o contexto dos dados, passam a ser retirados de um possível resultado.

Outra característica desse comparador e organizador semântico dos resultados está na estruturação das informações, em que os dados são reestruturados, de forma tal que isso auxilie no processo de verificação e estruturação, para serem mostrados, futuramente, para os usuários. O princípio aqui está em organizar os dados em uma estrutura que contemple o contexto dos dados, aderentes às ontologias utilizadas.

Assim, ao final desse processo, tem-se os dados organizados em uma nova estrutura, que segue a ontologia. Esses dados irão passar, então, por um processo de verificação da linguagem e da semântica, em que os resultados serão aplicados a algoritmos de aprendizagem de máquinas, para identificar se algumas das informações não têm o significado que se entende

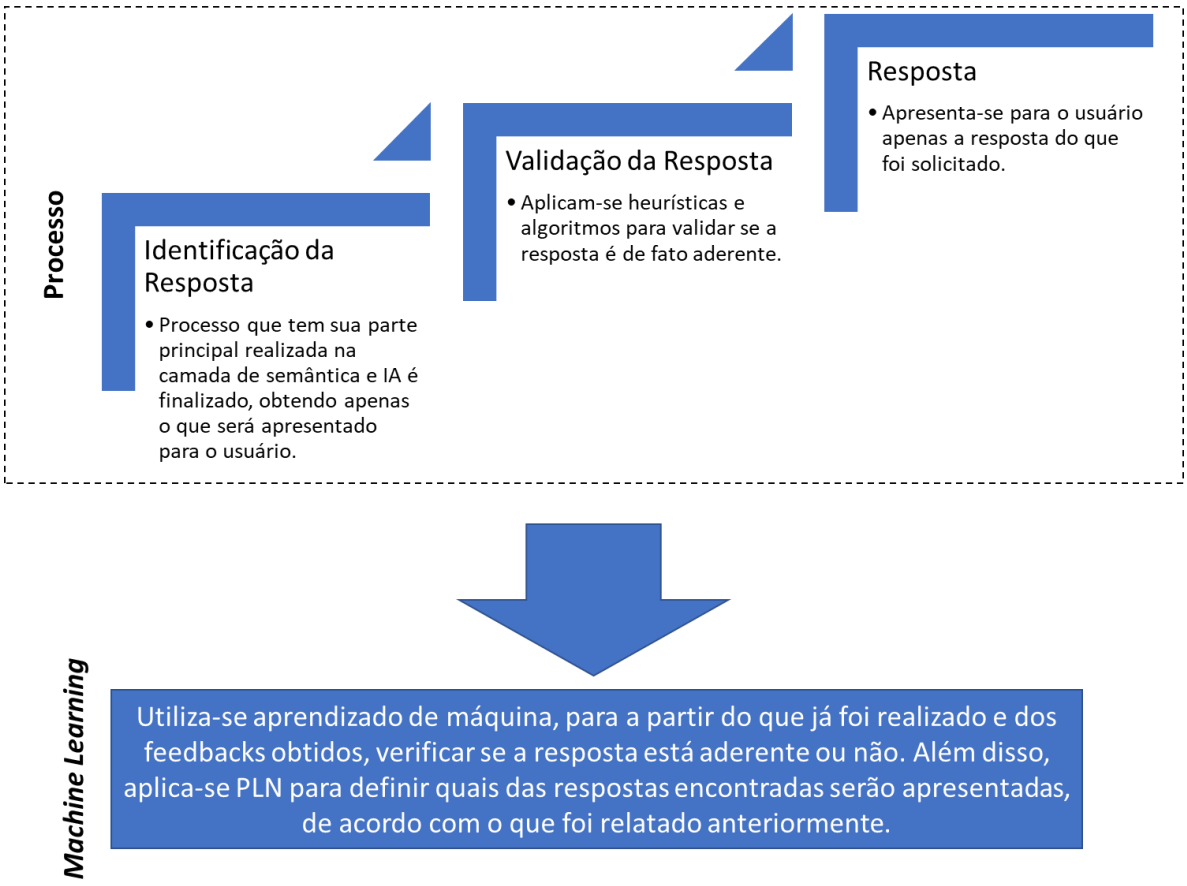
que elas têm. Esse processo está no contexto de *Natural Language Understanding*, em processos de classificação das informações, em que, a partir de uma grande base de textos treinados, é possível definir algumas classificações para as informações. Ao cruzar esses resultados com os dados organizados na estrutura da ontologia, tem-se um aumento ainda mais significativo no nível de semântica formal dos resultados.

Dessa forma, o final desse módulo do filtro, entrega as informações processadas, seguindo a estrutura do RDF, para que o criador da resposta possa validar e obter a melhor resposta a ser apresentada para o usuário.

#### 5.3.2.5 Criador da resposta

A última etapa do modelo é o criador de resposta, um módulo responsável por, a partir dos dados da resposta obtidos pelas camadas anteriores, apresentar os resultados para o usuário. Apresenta-se a seguir, na figura 38, o processo relativo ao criador das respostas.

**Figura 38** – Processo da resposta



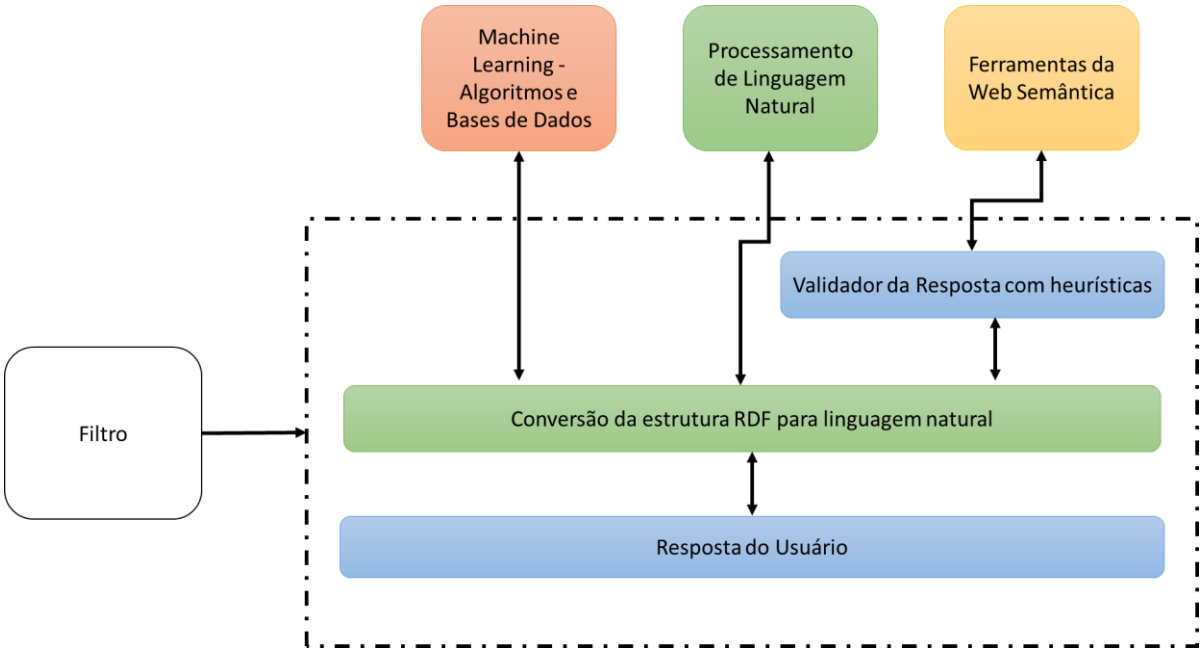

Fonte: elaborado pelo autor.

A partir do processo apresentado, verifica-se que, a partir dos resultados alcançados e tratados pelos módulos anteriores, são aplicados elementos que irão validar esses resultados e aprimorar aquilo que será apresentado ao usuário.

Destaca-se que serão utilizados algoritmos de *machine learning*, que são retroalimentáveis, visando a usar o aprendizado de máquina para melhorar os resultados apresentados aos usuários. Ressalta-se que a estrutura que vem da camada de semântica segue o formato RDF, devendo ser transformado em linguagem natural para que possa ser entendível pelo usuário, além de permitir novas questões, para sanar as suas necessidades informacionais.

Esse processo acontece no módulo criador de respostas. A arquitetura desse módulo pode ser visualizada na figura 39, que está vinculada ao modelo do projeto.

**Figura 39** – Arquitetura do criador de respostas




Fonte: elaborado pelo autor.

A arquitetura apresentada na figura 39 demonstra que, a partir dos resultados que estão organizados em formato de RDF, é realizado um processo de validação das respostas de acordo com as heurísticas que serão consideradas. Destaca-se que, aqui, se utilizam tanto os algoritmos de *machine learning*, que buscarão validar os resultados que foram obtidos, quanto a consulta a ontologias e às próprias bases do *Linked Data*, para validar se os resultados estão aderentes,

representados pelas ferramentas da Web Semântica na figura. Essas informações são trocadas entre esse validador e os algoritmos, para buscar validar se os resultados são apropriados.

Por fim, a partir do que foi validado, é convertido do formato RDF para linguagem natural, levando em consideração o tipo da pergunta que o usuário fez. Assim, tem-se a definição e a apresentação do resultado para os usuários.

### 5.3.3 Níveis de compreensão

Outro importante aspecto está na compreensão dos níveis de linguagem durante o processo, que vai se aprofundando no decorrer do modelo. Nesse sentido, esses níveis que estão relacionados com o processamento de linguagem natural são aprimorados quando estão vinculados com as Ferramentas da Web Semântica e com os algoritmos de *machine learning*.

A primeira questão relativa aos níveis de cada módulo está no aprofundamento de tais níveis de compreensão da linguagem conforme a sistemática do modelo ocorre. O primeiro nível faz referência à interação direta do usuário com o modelo, seja na pergunta ou na resposta, que está vinculada, em um primeiro momento, ao nível morfológico e léxico, pois é necessário que a fala ou a escrita seja interpretada a partir dos seus morfemas, para a sequência do modelo.

Posteriormente, os módulos identificador da pergunta e classificador da pergunta aprofundam o tratamento da linguagem realizando análises sintáticas e semânticas dos termos e das relações sintáticas existentes entre os termos da pergunta, além de realizar a compreensão do sentido dos termos. Essa compreensão que acontece tanto em um nível sintático, quanto semântico, é realizada pelo classificador da pergunta ao utilizar e compreender as relações entre os termos e a classificação de cada termo. Também nessa etapa, a definição de entidades acontece em um nível semântico, demonstrando o aprofundamento dos níveis com a evolução do modelo.

Por fim, na fase de enriquecimento semântico, tem-se o nível pragmático em que ocorre uma compreensão do contexto em que os termos estão inseridos, além de ter um aprofundamento na semântica dos termos. Isso é demonstrado pela criação de modelos conceituais criados a partir do entendimento dos termos pelas ontologias e pelas análises de PLN e *machine learning*.

Vale destacar que tais níveis são complementares e há uma relação entre eles nos diferentes módulos, conforme ocorre a sua evolução.

## 5.4 PROVA DE CONCEITO

Nesta subseção será apresentada uma prova de conceito, que demonstra como o modelo irá se comportar em um caso, simulando o funcionando de um sistema criado a partir da proposta desenvolvido nesta tese.

Anteriormente à demonstração de como o modelo se comporta, tem-se alguns elementos que precisam estar definidos e configurados para que o sistema possa funcionar de forma correta. Em especial, é fundamental definir ontologias que irão permitir a compreensão do domínio por parte do modelo, além de definir algumas ferramentas de *machine learning* que serão utilizadas e as fontes informacionais para realizar a recuperação da informação.

Vale destacar que a definição desses instrumentos está diretamente vinculada ao domínio escolhido para a realização dessa prova de conceito. Nesse caso, será utilizado o domínio de doenças humanas, por ter uma grande quantidade de ontologias e fontes informacionais que podem ser utilizadas para a realização dessa prova.

O quadro 10 apresenta os instrumentos e configurações definidas para a realização da prova de conceito a ser realizada no presente trabalho.

**Quadro 10** – Configurações e instrumentos utilizados pelo modelo na prova de conceito

| **Ferramenta/ Configuração** | **Descrição** | **Acesso** |
|---|---|---|
| Ontologia | Utiliza-se como ontologia principal a *Human Disease Ontology*, que apresenta diversas doenças classificadas quanto às categorias, ao tipo e à proliferação de doenças.<br>Exemplo de termo contido na ontologia é: *meningioma*, contendo diversas subcategorias e supercategorias, além de outras relações e classificações. | https://www.ebi.ac.uk/ols/ontologies/doid |
| Fontes informacionais | Uma das fontes informacionais utilizadas é o DisGeNET-RDF, uma base de doenças e genes em formatos de *Linked Data*. | https://www.disgenet.org/rdf |
| | Outra fonte informacional, mas que não é embasada em formato de *Linked Data* é o *MalaCards Human Disease Database*. Essa base contém diversas informações sobre doenças. | https://www.malacards.org/ |
| | Além dessas, outra fonte informacional que será utilizada é o Google Scholar, que providenciará textos que podem auxiliar na obtenção de informações acerca de doenças. | https://scholar.google.com.br/ |

Fonte: elaborado pelo autor.

O quadro 10 aponta uma ontologia, que tem destaque dentro do âmbito de doenças humanas, além de apresentar três fontes informacionais com características distintas que poderão contribuir para a obtenção de respostas para as perguntas dos usuários. Tem-se essas três bases, uma com foco no *Linked Data* (DisGeNET-RDF), outra como uma base de dados tradicional (*MalaCards Human Disease Database*) e uma terceira com o enfoque nos textos acadêmicos (Google Scholar).

Outro importante destaque está no uso de uma ferramenta de Inteligência Artificial e no processamento de linguagem natural que possibilitam a realização dessa prova de conceito. Essa definição busca demonstrar que as atuais ferramentas com esse enfoque são capazes de realizar o que este modelo busca. Para isso, utiliza-se como base o IBM Watson, uma plataforma de Inteligência Artificial da IBM, que contempla diversas funcionalidades para processamento de linguagem natural e aprendizagem de máquinas. O quadro 11 apresenta as funcionalidades que serão utilizadas para a realização dessa prova de conceito.

**Quadro 11** – Funcionalidades IBM Watson para a prova de conceito

| **Funcionalidade IBM Watson** | **Característica** | **Uso Modelo** |
|---|---|---|
| *Watson Natural Language Understanding* | Essa funcionalidade possibilita a classificação de termos, identificando conceitos, entidades e palavras-chave de um texto em linguagem natural. Além disso, essa funcionalidade é dotada de algoritmos de *machine learning* que, de acordo com o uso, possibilita o aprendizado de máquinas, uma vez que, ao ter novas palavras e novos conceitos, o sistema vai aprendendo e compreendendo a sua inserção dentro de cada cenário. | Utiliza para encontrar termos chaves na pergunta, além de ser utilizado nos textos que podem conter as respostas das perguntas. |
| *Watson Natural Language Classifier* | O classificador irá atuar para classificar qual é o sentido de uma determinada frase ou pergunta. A ideia dessa funcionalidade está em treinar o algoritmo com uma determinada base para compreender de que tema o usuário está tratando. Nesse sentido, é possível definir se o usuário está buscando informações de tratamento de uma doença ou se está buscando mais informações dessa doença, sendo um importante elemento para a classificação e o entendimento de perguntas. | Utiliza principalmente para classificar do que uma pergunta está se tratando. |
| *Watson Knowledge Studio* | Essa ferramenta é capaz de, partindo de um texto, compreender as diversas informações do domínio, classificando e definindo as relações existentes no texto. A possibilidade de realizar uma compreensão de um domínio, utilizando um modelo conceitual construído junto as ferramentas da Web Semântica, | Utiliza para análise de textos encontrados em fontes informacionais, favorecendo o |

| Funcionalidade IBM Watson | Característica | Uso Modelo |
|---|---|---|
| | torna a compreensão de um texto rápida, auxiliando o seu entendimento. | encontro da resposta. |
| *Watson Language Translator* | Esse módulo irá auxiliar na tradução de textos em português para inglês e vice-versa. Como muitas ontologias e textos que podem ter a resposta estão em inglês é necessário que essa ferramenta seja utilizada. | Tradução dos termos de inglês para português e vice-versa. |

Fonte: elaborado pelo autor.

O quadro 11 destaca quatro módulos do IBM Watson que podem ser utilizados para a realização da prova de conceito realizada neste trabalho. Esses quatro módulos demonstram que ferramentas atuais suportam e favorecem a criação de sistemas baseados no modelo proposto.

A partir dessas definições, inicia-se propriamente a prova de conceito, que partirá de uma determinada pergunta e dará uma resposta completa e que favoreça a recuperação da informação pelo usuário. Tem-se assim, a pergunta destacada no quadro 12.

**Quadro 12** – Pergunta da prova de conceito

| **Quais são os sintomas da doença meningioma do seio cavernoso?** |
|---|

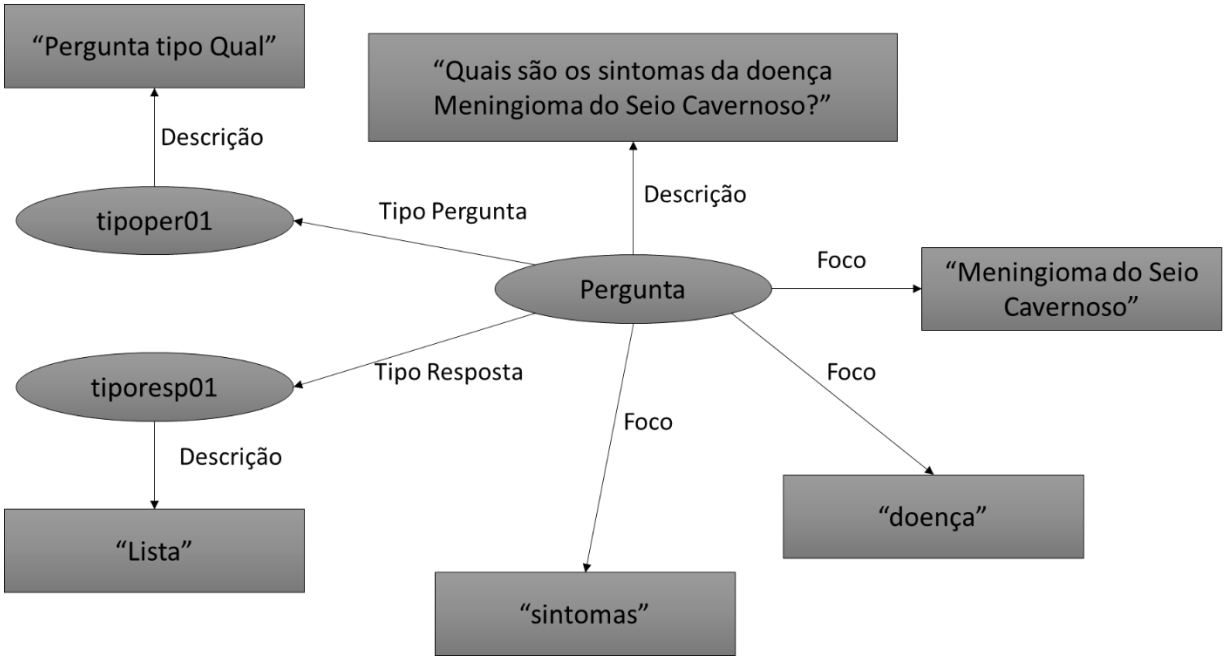
Fonte: elaborado pelo autor.

A pergunta realizada está dentro do contexto de doenças humanas e, no caso, o usuário busca identificar quais são os sintomas que a doença provoca. Nesse cenário, o primeiro aspecto a ser realizado no âmbito do modelo está na compreensão por parte do "identificador da pergunta", entendendo se o usuário fez a sua busca em formato de pergunta ou de palavras-chave. Assim, nessa primeira etapa, este módulo define que esses termos são relativos a uma pergunta a partir da estrutura sintática iniciada com termo tal como "Quais", e que utiliza a pontuação final "?".

Com essa definição de que é uma pergunta, a próxima etapa é realizada pelo módulo "classificador da pergunta", que irá transformá-la em uma estrutura em RDF, que reflete o tipo de pergunta efetuada pelo usuário, bem como um entendimento do que o usuário está buscando com a pergunta realizada.

Basicamente, essa classificação irá definir qual é o tipo da questão e o tipo da resposta, além de definir o foco da questão. No âmbito do modelo, a figura 40 define esses aspectos por meio de uma estrutura RDF.

**Figura 40** – RDF da classificação da pergunta




Fonte: elaborado pelo autor.

A figura 40 demonstra como a pergunta foi reestruturada, de modo que são entendidos que a pergunta se trata do tipo "Qual" e que o resultado deve ser uma lista. Além disso, definem-se os termos-chave, que são descobertos após a exclusão das palavras não relevantes, encontrando-se os termos: "sintomas", "doença" e "meningioma do seio cavernoso".

Tal estrutura é um modelo conceitual de RDF, demonstrando uma forma de criar uma estrutura que atenda à necessidade do modelo para a sua execução. Dessa forma, tem-se no modelo relações que vinculam a pergunta ao seu foco, seu tipo de pergunta e seu tipo de resposta.

Após essa etapa, os termos definidos como foco serão enriquecidos no módulo "enriquecedor semântico dos termos". Essa etapa encontrará informações relevantes do domínio, que estão vinculadas ao que o usuário está buscando, utilizando ontologias e algoritmos de entendimento dos termos com *machine learning*.

O primeiro aspecto trata da expansão e enriquecimento comparando com a ontologia, que dará informações relevantes sobre os termos em questão. Após isso, esse mesmo trabalho

é realizado com o uso de ferramentas de Inteligência Artificial e processamento de linguagem natural e, ao final, haverá um modelo em RDF expandido a partir dos termos buscados.

O quadro 13 mostra algumas informações encontradas na ontologia e nos algoritmos de PLN que foram base para a expansão da busca. Vale destacar que, anterior a esse processo, ocorre a tradução do termo para o inglês para que a sua comparação seja mais efetiva, de modo que se busca e se compara tanto em português, quanto em inglês. Para fazer essa busca, utiliza-se o motor de busca em SPARQL, que possibilita a expansão com o uso de ontologias, e para *machine learning* utilizam-se as ferramentas do IBM Watson.

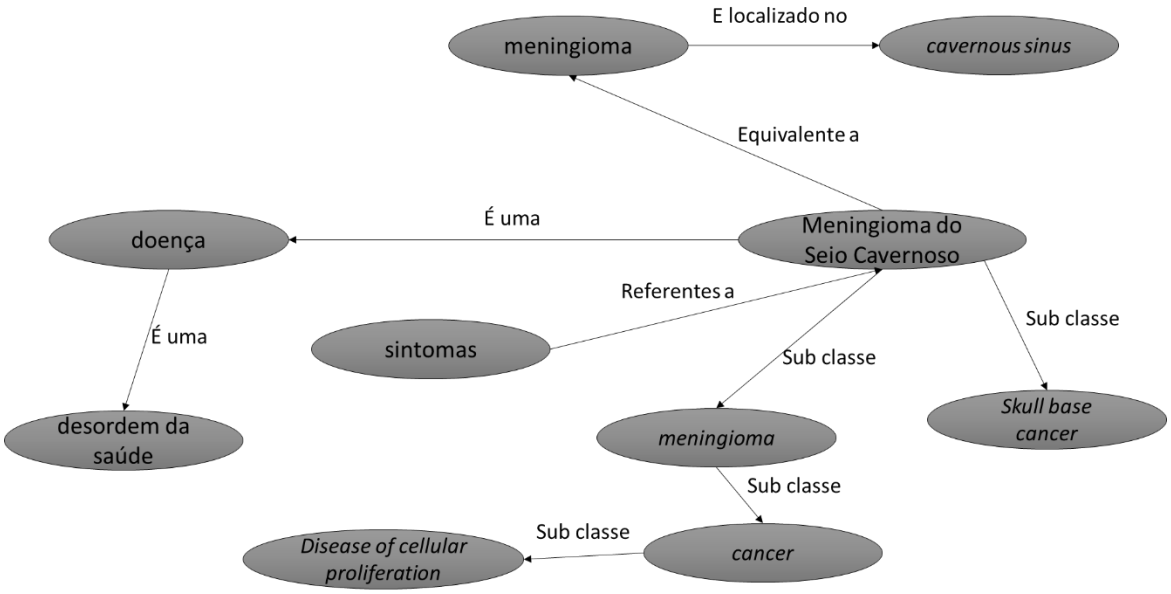
**Quadro 13** – Expansão dos termos com a ontologia e *machine learning*

| **Termo PT** | **Termo EN** | **Ontologia** | ***Machine Learning*** |
|---|---|---|---|
| Sintomas | *Symptoms* | Não encontrado | Compreensão de que sintomas faz referência ao meningioma do seio cavernoso. |
| Doença | *Disease* | Termo superclasse na ontologia. Código DOID:4 | Trata-se de uma desordem relativa à saúde. |
| Meningioma do seio cavernoso | *Cavernous Sinus Meningioma* | Termo subclasse de meningioma e de *skull base cancer*, Relacionamento de equivalência com “meningioma”, localizado no *cavernous sinus*, além de outros vínculos encontrados. | Compreensão de que diz respeito a algo relativo a Cavernous sinus. |

Fonte: elaborado pelo autor.

Partindo das informações obtidas com a ontologia e a análise com o auxílio de *machine learning* é possível refazer um grafo em RDF que relaciona essas informações e permite uma expansão quanto aos termos que representam a busca do usuário. Assim, a figura 41 demonstra o grafo construído com as informações obtidas. Vale destacar que as informações contidas na figura são representações conceituais da estrutura que o RDF teria, caso fosse estruturado, com seus relacionamentos e classes.

**Figura 41** – Grafo com o enriquecimento dos termos da pergunta




Fonte: elaborado pelo autor.

Com a figura 41, é possível verificar que os três termos iniciais foram expandidos, tanto relacionando e interligando os termos entre si, quanto inserindo novas informações que podem auxiliar no processo de busca. Todas essas novas informações foram obtidas a partir das análises nas ontologias e nos algoritmos de *machine learning*.

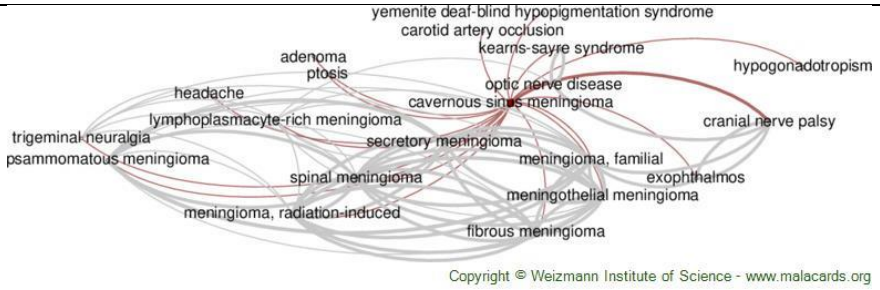
Após esse processo, com os termos enriquecidos, inicia-se o processo de busca, utilizando-se as fontes informacionais definidas anteriormente. Nesse sentido, há duas principais formas de realizar a busca: a primeira utilizando a estrutura semântica do RDF para encontrar informações no *Linked Data* e a segunda utilizando uma expressão de busca tradicional para encontrar informações em fontes de bases de dados tradicionais ou textuais.

O quadro 14 demonstra alguns resultados encontrados nas três fontes informacionais que foram utilizadas para a recuperação da informação. Destaca-se que cada fonte foi consultada com processos de recuperação distintos, sendo o uso do SPARQL[2] para o DisGeNET-RDF, que é embasado nos princípios do *Linked Data*, o uso de API para o Google Scholar[3] e o processo de extração da consulta no website para o *MalaCards Human Disease Database*[4], sendo considerado que o enriquecimento realizado deveu-se à ontologia e aos algoritmos de *machine learning*.

[2] Uso do SPARQL Endpoint do DisGeNET-RDF. Disponível em: <http://rdf.disgenet.org/sparql/>. Acesso em: 26 fev. 2020.
[3] Uso do Scholarly 0.25, biblioteca da linguagem Python. Disponível em : <https://pypi.org/project/scholarly/>. Acesso em: 26 fev. 2020.
[4] Extração a partir da busca realizada em: <https://www.malacards.org/>. Acesso em: 26 fev. 2020.

**Quadro 14** – Exemplos de resultados obtidos nas fontes informacionais

| **Fontes Informacionais** | **Informações encontradas** |
|---|---|
| DisGeNET-RDF | Name: Meningioma<br>UMLS CUI: C0025286<br>Type: disease<br>MeSH Class: Neoplasms; Nervous System Diseases<br>MeSH: D008579<br>OMIM: 190040;601728<br>Semantic Type: Neoplastic Process<br>Phenotypic abnormality: Abnormality of the nervous system; Neoplasm<br>Disease Ontology: disease of anatomical entity; disease of cellular proliferation |
| | Name: Malignant Bone Neoplasm<br>UMLS CUI: C0279530<br>Type: disease<br>MeSH Class: Musculoskeletal Diseases; Neoplasms<br>MeSH: D001859<br>OMIM: None<br>Semantic Type: Neoplastic Process<br>Phenotypic abnormality: None<br>Disease Ontology: disease of anatomical entity; disease of cellular proliferation |
| | Name: Skull Base Meningioma<br>UMLS CUI: C1335976<br>Type: disease<br>MeSH Class: None<br>MeSH: None<br>OMIM: None<br>Semantic Type: Neoplastic Process<br>Phenotypic abnormality: None<br>Disease Ontology: disease of anatomical entity; disease of cellular proliferation |
| *MalaCards Human Disease Database* | Classifications: MalaCards categories:<br>Global: Cancer diseases<br>Anatomical: Neuronal diseases bone diseases<br>See all MalaCards categories (disease lists)<br><br>Disease Ontology DOID:4435<br><br>Cavernous Sinus Meningioma, also known as meningioma of the cavernous sinus, is related to cranial nerve palsy and meningioma, radiation-induced, and has symptoms including seizures and headache. An important gene associated with Cavernous Sinus Meningioma is DHFR (Dihydrofolate Reductase), and among its related pathways/superpathways is NAD metabolism. Affiliated tissues include bone, brain and pituitary, and related phenotype is increased viability with MLN4924 (a NAE inhibitor).<br><br>Graphical network of the top 20 diseases related to Cavernous Sinus Meningioma: |

| **Fontes Informacionais** | **Informações encontradas** |
|---|---|
| |  <br><br>UMLS symptoms related to Cavernous Sinus Meningioma:<br>seizures, headache<br>GenomeRNAi Phenotypes related to Cavernous Sinus Meningioma according to GeneCards Suite gene sharing:<br>Increased viability with MLN4924 (a NAE inhibitor) - GR00250-A-3<br><br>... |
| Google Scholar | 'bib': {'abstract': 'Intracranial meningiomas are known to infiltrate ' 'surrounding structures such as the calvaria and dural ' 'sinuses, and the brain itself. The issue of whether ' 'meningiomas invade major intracranial arteries is of ' 'clinical importance, particularly in the case of ' 'meningiomas of …',<br>'author': 'MJ Kotapka and KK Kalia and AJ Martinez and LN Sekhar',<br>'eprint': 'https://pdfs.semanticscholar.org/e273/662839f7ffee8efc3d4316af85a97d8c703d.pdf',<br>'title': 'Infiltration of the carotid artery by cavernous sinus ' 'meningioma',<br>'url': 'https://thejns.org/view/journals/j-neurosurg/81/2/article-p252.xml'} |
| | 'bib': {'abstract': 'Object. The authors sought to assess the functional ' 'tolerance and tumor control rate of cavernous sinus ' 'meningiomas treated by gamma knife radiosurgery (GKS). ' 'Methods. Between July 1992 and October 1998, 92 patients ' 'harboring benign cavernous sinus …',<br>'author': 'PH Roche and J Régis and H Dufour and HD Fournier…',<br>'title': 'Gamma knife radiosurgery in the management of cavernous ' 'sinus meningiomas',<br>'url': 'https://thejns.org/view/journals/j-neurosurg/93/supplement_3/article-p68.xml'} |
| | 'bib': {'abstract': 'Background Fractionated stereotactic radiotherapy (FSRT) ' 'combines the precision of stereotactic positioning with ' 'the radiobiologic advantage of dose fractionation. ' 'Methods From June 1997 to June 2001, 30 patients with ' 'cavernous sinus meningiomas were treated …',<br>'author': 'M Brell and S Villà and P Teixidor and A Lucas and E ' 'Ferrán and S Marín…', |

| **Fontes Informacionais** | **Informações encontradas** |
|---|---|
| | 'title': 'Fractionated stereotactic radiotherapy in the treatment of ' 'exclusive cavernous sinus meningioma: functional outcome, ' 'local control, and tolerance', 'url': 'https://www.sciencedirect.com/science/article/pii/S0090301905004404'} |
| | 'bib': {'abstract': 'Introduction: The purpose of this study was to evaluate ' 'the efficacy and safety of stereotactic radiosurgery as ' 'primary management for patients with imaging defined ' 'cavernous sinus meningiomas. Methods: Between 1992 and ' '2001, 49 patients had radiosurgery for dural …', 'author': 'BE Pollock and SL Stafford', 'title': 'Results of stereotactic radiosurgery for patients with ' 'imaging defined cavernous sinus meningiomas', 'url': 'https://www.sciencedirect.com/science/article/pii/S0360301605000404' |

Fonte: elaborado pelo autor.

Os resultados apresentados no quadro 14 são oriundos de diversas fontes, com mais de um resultado obtido de cada fonte, possibilitando, assim, que os resultados sejam explorados para se atingir o que o usuário espera como resposta.

Após a obtenção dos resultados, a próxima etapa é o "validador" que irá verificar se os resultados obtidos estão ou não aderentes ao que o usuário está buscando. Para isso, são verificados se os resultados estão aderentes ao contexto, bem como se há uma relação entre o resultado e o que o usuário está buscando.

O processo realizado pelo validador é apresentado no quadro 15, que demonstra se os resultados estão ou não aderentes ao contexto e ao que o usuário está esperando. No quadro, a primeira coluna apresenta a fonte, a segunda demonstra o número do resultado de acordo com o quadro 14, a terceira coluna apresenta se aquele resultado está ou não aderente ao contexto e a quarta coluna apresenta se o resultado está ou não aderente ao que o usuário está buscando como resposta.

**Quadro 15** – Processo do validador para verificar aderência dos resultados

| **Fonte** | **# Resultado** | **Aderência ao contexto** | **Aderência à resposta** |
|---|---|---|---|
| DisGeNET-RDF | 1 | Sim | Sim |
| | 2 | Sim | Não |
| | 3 | Sim | Não |
| MalaCards Human Disease Database | 1 | Sim | Sim |
| Google Scholar | 1 | Sim | Não |
| | 2 | Sim | Não |
| | 3 | Sim | Sim |

| Fonte | # Resultado | Aderência ao contexto | Aderência à resposta |
|---|---|---|---|
| | 4 | Sim | Não |

Fonte: elaborado pelo autor.

O quadro 15 demonstra que houve várias respostas, três, que estão aderentes ao contexto e às respostas esperadas pelos usuários. Dessa forma, a próxima etapa no modelo é o "filtro", responsável por definir e organizar os resultados encontrados, comparando com as estruturas semânticas utilizadas, além de utilizar os algoritmos de *machine learning*.

Nessa etapa, apenas os resultados que estão aderentes ao contexto e às respostas são considerados para a sequência do processo. Assim, obtêm-se as partes de cada resposta que tem a resposta esperada pelo usuário, para que se possa organizar e criar uma estrutura que vise a uma resposta completa e satisfatória.

O quadro 16 apresenta os resultados obtidos a partir de cada fonte.

**Quadro 16** – Parte do resultado com a resposta da pergunta

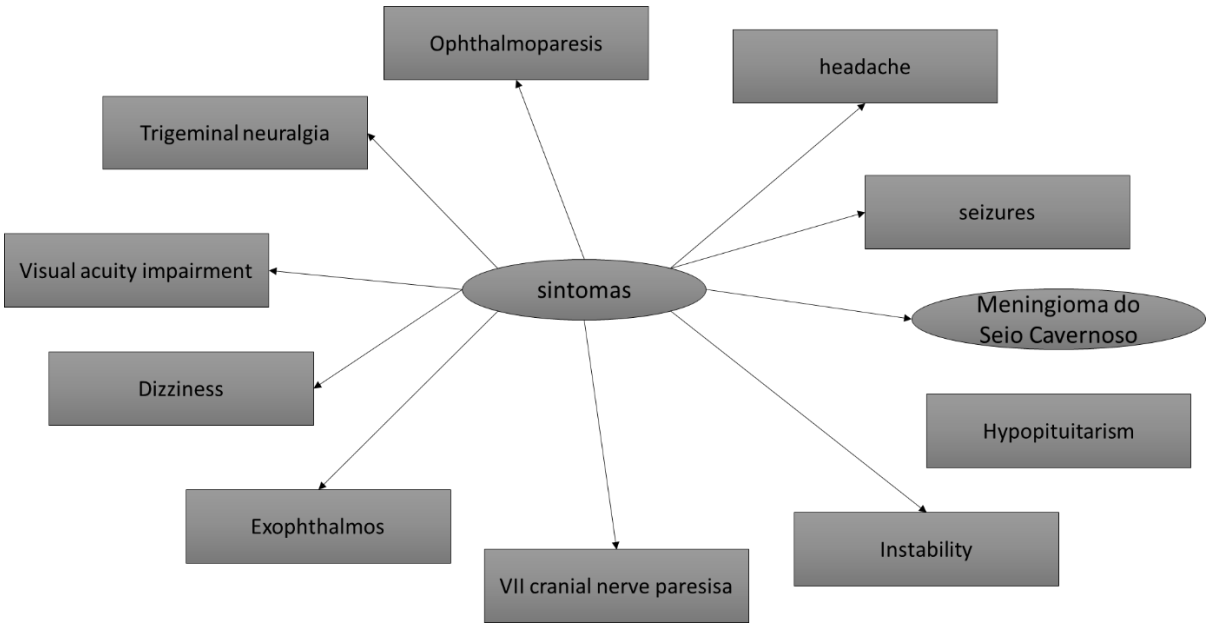
<table>
<tr><th>Fonte</th><th>Parte da resposta com resultado</th></tr>
<tr><td>DisGeNET-RDF</td><td>Phenotypic abnormality: Abnormality of the nervous system; Neoplasm</td></tr>
<tr><td>MalaCards Human Disease Database</td><td>UMLS symptoms related to Cavernous Sinus Meningioma:<br>seizures, headache</td></tr>
<tr><td>Google Scholar</td><td>
<table>
<tr><th>Symptoms</th><th>Patients (n [%])</th></tr>
<tr><td>Headache</td><td>6 (20%)</td></tr>
<tr><td>Ophthalmoparesis</td><td>17 (56.6%)</td></tr>
<tr><td>Trigeminal neuralgia</td><td>12 (40%)</td></tr>
<tr><td>Visual acuity impairment</td><td>14 (46.6%)</td></tr>
<tr><td>Dizziness</td><td>1 (3.3%)</td></tr>
<tr><td>Exophthalmos</td><td>4 (13.3%)</td></tr>
<tr><td>VII cranial nerve paresisa</td><td>3 (10%)</td></tr>
<tr><td>Instability</td><td>3 (10%)</td></tr>
<tr><td>Hypopituitarism</td><td>1 (3.3%)</td></tr>
<tr><td>Epileptic seizures</td><td>1 (3.3%)</td></tr>
</table>
</td></tr>
</table>

Fonte: elaborado pelo autor.

No quadro 16, é possível visualizar que as três fontes contemplaram resultados possíveis de serem apresentados para os usuários e que podem ser estruturados de uma forma mais organizada, considerando aspectos semânticos. Vale destacar que, para a obtenção desses resultados, é necessário analisar os textos e as suas estruturas para encontrar as respostas relevantes dentro dos diferentes documentos.

Partindo dos resultados, constrói-se uma estrutura semântica de acordo com o contexto, que irá relacionar os diferentes resultados, avaliando o contexto em que as partes se vinculam. A figura 42 apresenta o grafo criado a partir dessas respostas.

**Figura 42** – Estrutura semântica construída a partir das respostas obtidas




Fonte: elaborado pelo autor.

A figura 42 apresenta a estrutura semântica, contendo as respostas para a pergunta realizada pelo usuário. Com essa estrutura criada, a próxima etapa, relativa ao módulo “criador da resposta”, pode ser realizada.

Nessa etapa, acontece a conversão da estrutura RDF, que contém os resultados em linguagem natural a ser apresentada ao usuário. A realização desse processo é facilitada pela estrutura, mas demanda o uso de processos de PLN, *machine learning* e das ferramentas da Web Semântica.

Após a tradução dos resultados obtidos, a frase que contempla a resposta da pergunta do usuário é criada, conforma apresentado no quadro 17.

**Quadro 17** – Resposta dada ao usuário

**Os sintomas da doença meningioma do seio cavernoso são: convulsão[1] [2], dor de cabeça[1] [2], Oftalmoparesia[1], neuralgia trigeminal[1], comprometimento da acuidade visual[1], tontura[1], exoftalmia[1], paresia do nervo craniano[1], instabilidade[1] e hipopituitarismo[1].**

**Fontes:**
[1] Fractionated stereotactic radiotherapy in the treatment of exclusive cavernous sinus meningioma: functional outcome, local control, and tolerance. Marta Brell, Salvador Villa, Pilar Teixidor, Anna Lucas, Enric Ferrán, Susanna Marín e Juan Jose Acebes. Disponível em: https://www.sciencedirect.com/science/article/pii/S0090301905004404
[2] MalaCards Human Disease Database. Disponível em: https://www.malacards.org/card/cavernous_sinus_meningioma?search=Cavernous%20Sinus%20Meningioma

Fonte: elaborado pelo autor.

A partir da frase do quadro 17, verifica-se que essa resposta, que já foi avaliada e definida como a correta, utilizou os diversos elementos que foram obtidos a partir das ontologias, das fontes informacionais e da compreensão da linguagem natural para ser montada. As diversas relações definidas, como a relação dos sintomas com a doença, o entendimento de que o meningioma do seio cavernoso era uma doença, a união entre os diversos sintomas obtidos de distintas fontes informacionais, tornaram o processo complexo, tendo exigido a união entre as ferramentas da Web Semântica, os algoritmos de *machine learning* e as técnicas de processamento de linguagem natural.

A prova de conceito realizada passou por todas as etapas do modelo, demonstrando o seu funcionamento, atendendo a recuperação da informação com o uso de linguagem natural e com Inteligência Artificial.

# 6 CONSIDERAÇÕES FINAIS

Este trabalho propõe um modelo de recuperação da informação que aproxima e relaciona este campo de estudos com a Inteligência Artificial, processamento de linguagem natural e Web Semântica, tendo a perspectiva da Ciência da Informação. Dessa forma, este modelo é capaz de criar um meio da recuperação com o enfoque no usuário, de maneira que ele possa ter um acesso facilitado para a informação, uma vez que o processo de recuperação acontece com o uso de linguagem natural e considera a semântica dos termos e da busca do usuário.

Adicionalmente, aponta-se que este trabalho avança nas possibilidades que o campo de estudo da Web Semântica possui, uma vez que realiza uma aproximação necessária entre os estudos e as ferramentas da Web Semântica com a área da Inteligência Artificial, em especial de processamento de linguagem natural.

Nesse contexto, a revisão teórica desenvolvida demonstra como quatro conceitos-chave, recuperação da informação, Web Semântica, Inteligência Artificial e processamento de linguagem natural são apresentados pela comunidade acadêmica, destacando a sua evolução nas últimas décadas, mas, em especial, como se encontram no presente momento e a possibilidade de um relacionamento efetivo entre essas quatro áreas.

Essa primeira etapa demonstrou uma série de interseções possíveis entre esses campos, em especial no que tange à necessidade de aprimorar o nível de semântica formal para que a recuperação da informação seja aprimorada. Sobretudo as ferramentas da Web Semântica, a recuperação da informação e as técnicas de processamento de linguagem natural podem contribuir para que exista uma melhor compreensão para o sistema de um texto e de uma base de dados.

A partir desse cenário, discutiu-se o modelo de recuperação da informação e, em destaque, o do *Question Answering* como um modo mais natural para os usuários encontrarem informações que estejam buscando. Isso ocorre, pois, a utilização dos princípios de processamento de linguagem natural para os usuários buscarem informações por meio de perguntas e respostas, bem como a adoção da Inteligência Artificial e do *Question Answering* são capazes de transformar o modo como sistemas de recuperação de informação satisfazem as necessidades informacionais dos usuários.

Assim, o modelo proposto por esse trabalho parte da Inteligência Artificial e do processamento de linguagem natural, para utilizar esse benefício que o *Question Answering* possui, além de inserir uma série de ferramentas da Web Semântica, nas diversas etapas do processo. Realiza-se uma aproximação entre campos de estudos distintos, Web Semântica,

Inteligência Artificial e processamento de linguagem natural, mas que possuem objetivos semelhantes, na tentativa de tornar o processo de busca mais eficiente, ao considerar o significado e o contexto dos usuários e das fontes informacionais.

Trazendo o objetivo geral deste trabalho, que está em propor um novo processo de recuperação da informação, que redesenha esse campo de estudos, a partir da aproximação da linguagem computacional com a linguagem natural, utilizando os princípios da representação da informação para que o significado e o contexto dos dados estejam explícitos para o processo da busca, aponta-se que tal objetivo foi alcançado a partir do modelo proposto e da discussão realizada para a construção e definição desse modelo, além de atender o conjunto dos objetivos específicos relatados. Além disso, ao apresentar a prova de conceito, demonstra-se como o modelo se comporta, destacando que o modelo propicia esse novo processo de recuperação da informação.

Primeiramente, o objetivo específico de propor um modelo conceitual que abarque os processos da recuperação da informação centrado no uso de ferramentas da Web Semântica, foi atingido no capítulo que apresenta o modelo de recuperação da informação, quando, a partir da seção de recuperação da informação, Inteligência Artificial e processamento de linguagem natural, define-se o modelo criado.

Em seguida, o objetivo específico de conceituar e relacionar a Ciência da Informação, a recuperação da informação, a Web Semântica, a Inteligência Artificial e o processamento de linguagem natural, foi atingido na parte teórica, que apresenta os diversos conceitos trabalhados, bem como exploram-se as suas ligações e vínculos a partir da perspectiva do trabalho.

O terceiro objetivo específico, que concerne à apresentação e à discussão de como a Inteligência Artificial pode apoiar a recuperação da informação, é alcançado quando é apresentado o uso da Inteligência Artificial no âmbito da recuperação da informação. Além disso, na subseção de recuperação da informação e processamento de linguagem natural, apresenta-se essa relação e aprofunda-se o entendimento dos impactos da IA nos processos de recuperação.

O quarto objetivo é relativo à definição de como as ontologias poderiam ser utilizadas para a contextualização dos dados nas fases da recuperação da informação, sendo atingido na definição do modelo, que demonstra como as ontologias são utilizadas nas diversas etapas da recuperação da informação.

Por fim, o quinto objetivo específico trata da utilização do *Linked Data* como fonte de informação para auxiliar o processo de expansão da busca, que foi alcançado na própria

definição do modelo, quando foi utilizado o *Linked Data* tanto como fonte de informação, quanto para auxiliar na contextualização dos dados, junto com outros instrumentos da Web Semântica.

Com todos os objetivos atingidos e com a prova de conceito realizada, o presente trabalho traz uma contribuição à área de Ciência da Informação, no que tange a uma evolução dos processos de recuperação da informação. A integração de Inteligência Artificial junto com a Web Semântica nesse processo é capaz de inserir a recuperação da informação, quando analisada e compreendida no âmbito da Ciência da Informação, em um novo cenário, capaz de tornar o processo mais semântico, inteligente e próximo das necessidades dos usuários.

Nesse aspecto, o acesso à informação pode ser ampliado quando o uso de sistemas de recuperação da informação ocorre com o apoio de linguagem natural, sem a necessidade da inserção de palavras-chave ou operadores booleanos. Assim, o acesso à informação se torna muito mais fácil e democrático, visto que o usuário não necessita ter um conhecimento prévio das técnicas de buscas para realizar uma pesquisa.

Adicionalmente, aponta-se que a Inteligência Artificial, quando compreendida de forma conceitual e enquanto um campo de estudos vinculado à Ciência da Computação, é capaz de auxiliar a Ciência da Informação a uma evolução, em que as tecnologias computacionais permitem um aprimoramento e a criação de novos campos de estudos. Dessa forma, a Inteligência Artificial pode ser redescoberta, rediscutida e evoluir como um campo de estudos vinculado a Ciência da Informação. Este trabalho contribui para isso, trazendo a Inteligência Artificial para ser analisada dentro do contexto epistemológico e teórico da Ciência da Informação, compreendendo o seu papel como uma ciência interdisciplinar.

Outro importante elemento que é parte central do trabalho desenvolvido diz respeito à Web Semântica, por meio de seus conceitos, tecnologias, ferramentas e aplicações. A Web Semântica foi utilizada como o elemento que permitiu a compreensão dos termos, levando em consideração a semântica, aspecto que é trabalhado no âmbito do processamento de linguagem natural, mas que é expandido e aprimorado quando está vinculado às ferramentas da Web Semântica. A partir da Inteligência Artificial e do processamento de linguagem natural, a Web Semântica pode ser explorada com mais profundidade, podendo efetivamente diminuir a lacuna entre os elementos computacionais e a linguagem humana.

Destaca-se que este trabalho se vincula e demonstra a viabilidade da proposta inicial de Tim Berners-Lee, James Hendler e Ora Lassila (2001) para a Web Semântica, ao indicar caminhos e possibilidades de aproximar e interligar as linguagens computacional e natural. Desta feita, o modelo proposto, junto a sua prova de conceito demonstram que a expressão do

significado, de modo que computador consiga entender o que uma pessoa precisa, de forma natural, é possível com as atuais tecnologias.

Neste sentido, há uma aproximação do processo de recuperação da informação com o usuário, trazendo-o para o centro do processo e permitindo que o modo como o usuário encontra a informação possa ser mais natural, com maior significado e mais expressivo. Assim, o processo de recuperação da informação é expandido, demonstrando que a apresentação de links não é aderente à linguagem natural do usuário, e com a proposição deste modelo, é possível tornar o processo mais natural e aprimorado, utilizando aspectos humanos no processo de recuperação.

Portanto, com este trabalho, foi possível validar e definir um novo meio de recuperação da informação, que parte dos conceitos e dos elementos da Ciência da Informação, apoiado por Inteligência Artificial, processamento de linguagem natural e Web Semântica, permitindo assim um avanço e uma evolução nesses campos de estudos.

Vale destacar que a partir dessa nova visão da recuperação da informação, realiza-se uma mudança no modo como a Ciência da Informação se relaciona com a recuperação da informação, partindo da adoção de conceitos e tecnologias da Inteligência Artificial nesse processo.

Aponta-se ainda que a relação entre Ciência da Informação e Ciência da Computação guiou o desenvolvimento desta pesquisa, se complementando e utilizando elementos que permitiram a definição do modelo proposto. O aspecto técnico da Ciência da Computação ao demonstrar aplicações e teorias da Inteligência Artificial, junto as discussões e aplicações da Ciência da Informação, por meio da recuperação da informação, foram expandidos quando unido as teorias e práticas da Web Semântica, possibilitando a interdisciplinaridade base do trabalho.

Aponta-se algumas limitações do presente trabalho, como a necessidade de se utilizar ontologias e outras ferramentas da Web Semântica, que podem inviabilizar a aplicação do modelo caso não sejam construídos tais elementos. Este aspecto pode ser corrigido com o avanço de pesquisas que buscam automatizar o processo de criação de ontologias, tornando assim, o modelo proposto mais flexível e adequado para um maior número de ambientes informacionais digitais.

Enquanto trabalhos futuros, busca-se realizar a implementação do modelo, utilizando como base algum cenário, visando a construir um sistema de recuperação da informação embasado no modelo criado. Adicionalmente, pode-se realizar um trabalho que demonstra qual o impacto no nível de semântica formal alcançado, e a consequência nos resultados obtidos,

quando se utiliza apenas fontes informacionais baseadas no formato de *linked data*, ou apenas fontes informacionais textuais. Esse trabalho pode demonstrar a eficácia da utilização, enquanto fonte para o processo de recuperação da informação, do *linked data*.

Além disso, é possível aplicar o modelo proposto em diversos cenários, criando mecanismos de recuperação da informação que possam ser utilizados para aprimorar sistemas de informação especialistas. Um outro cenário seria a aplicação em repositórios digitais, para aprimorar o atendimento às necessidades informacionais dos usuários nestes ambientes, trazendo a linguagem natural como meio da interface entre o usuário e as plataformas.

Outro trabalho futuro está na exploração da relação entre o modelo proposto neste trabalho e a Encontrabilidade da Informação, termo proposto por Vechiato e Vidotti (2014), visto que tanto o modelo quanto o conceito proposto pelos autores se vinculam à recuperação da informação e podem se complementar para aprimorar o modo como o usuário busca e encontra as informações nos ambientes informacionais digitais.

# REFERÊNCIAS


ABDI, A; IDRIS, N; AHMAD, Z. QAPD: an ontology-based question answering system in the physics domain. **Soft Computing**, v. 22, n. 1, p. 213-230, 2018.

ALLAM, A. M. N.; HAGGAG, M. H. The *Question Answering* systems: A survey. **International Journal of Research and Reviews in Information Sciences**, v. 2 n. 3 p. 211–220. 2012. Disponível em: <https://www.researchgate.net/profile/Ali_Allam/publication/281969283_The_Question_Answering_Systems_A_Survey/links/5600003308ae07629e51fe42.pdf>. Acesso em: 25 set. 2017.

ALMANSA, L. F. **Um framework de *Question Answering* nos domínios de epigenética e de imagens citológicas de tireoide**. 2016. 88f. Dissertação (Mestrado em Ciências) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto. 2016.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) -Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

ANDREATA, G. H. S. **O uso de processamento de linguagem natural para a análise de sentimentos na rede social Reddit**. 2018. Disponível em: <https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3804/TCC%20Guilherme%20Henrique%20Santos%20Andreata.pdf?sequence=1>

ARANTES, L. O. **Documentação semântica no apoio à integração de dados e rastreabilidade.** 2010. Dissertação de Mestrado. Disponível em: <http://repositorio.ufes.br/jspui/bitstream/10/6396/1/Dissertacao%20Lucas%20de%20Oliveira%20Arantes.pdf>. Acesso em: 14 maio 2020.

ASIAEE, A. H. et al. A framework for ontology-based *Question Answering* with application to parasite immunology. **Journal of biomedical semantics**, v. 6, n. 1, p. 1, 2015.

BAEZA-YATES; R.; RIBEIRO-NETO, B. **Modern information retrieval.** New York: ACM Press, 1999.

BARROW, H. G.; TENENBAUM, J. M. Computational vision. **Proceedings of the IEEE**, v. 69, n. 5, p. 572-595, 1981. Disponível em: <https://ieeexplore.ieee.org/abstract/document/1456294>. Acesso em: 31 dez. 2019.

BECHHOFER, S. et al. **OWL Web Ontology Language Reference**. 2004. Disponível em: < http://www.w3.org/TR/owl-ref/>. Acesso em: 08 abr. 2020.

BECKETT, D. et al. **RDF 1.1 Turtle**. 2014. Disponível em: https://www.w3.org/TR/turtle/>. Acesso em: 03 jul. 2019.

BELLMAN, R**. An introduction to artificial intelligence**: Can computers think?. Thomson Course Technology, 1978.

BERNARDO, R. O. SANTACHÉ, A.. **Aplicação de chatbots no desenvolvimento de jogos em saúde.** UNICAMP. 2017. Disponível em: <http://www.ic.unicamp.br/~reltech/PFG/2017/PFG-17-22.pdf>. Acesso em: 27 dez. 2019.

BERNERS-LEE, T. et al. **Uniform Resource Identifier (URI)**: Generic Syntax. 2005. Disponível em: <https://tools.ietf.org/html/rfc3986>. Acesso em: 17 jun, 2020.

BERNERS-LEE, T. **Information management**: a proposal. 1989. Disponível em: <https://www.w3.org/History/1989/proposal.html>. Acesso em> 09 dez. 2019.

BERNERS-LEE, T. **Linked Data principles.** 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 05 jul. 2019.

BERNERS-LEE, T. ***Linked Data*** **principles**. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html> Acesso em: 25 set. 2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

BERNERS-LEE, T.Semantic **Web road map**. 1998. Disponível em:<http://www.w3.org/DesignIssues/Semantic.html>. Acesso em: 18 set. 2005.

BERNERS-LEE, T.; CAILLIAU, R. **WorldWideWeb**: Proposal for a HyperText project, 1990. Disponível em: <https://www.w3.org/Proposal.html>. Acesso em: 09 jan. 2020.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked Data**: the story so far. International Journal on Semantic Web and Information Systems, v. 5, n. 3, p. 1-22, 2009. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 05 jul. 2019.

BORKO, H. Information science: what is it? **American documentation** 19.1, 3-5, 1968.

BORLUND, P. The concept of relevance in IR. **Journal of the American Society for Information Science and Technology**, v. 54, n. 10, p. 913–925, ago. 2003.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. 1997. 227 f. Tese (Doutorado)-Centre for Telematics for Information Technology, University of Twence, Enschede, 1997. Disponível em: <http://doc.utwente.nl/17864/1/t0000004.pdf>. Acesso em: 25 set. 2019

BOUZIANE, A. et al. TOWARD AN ARABIC QUESTION ANSWERING SYSTEM OVER LINKED DATA. **Jordanian Journal of Computers and Information Technology (JJCIT**), v. 4, n. 02, 2018.

BRAY, T. et al. Extensible markup language (XML) 1.1 (Second Edition). **World Wide Web Consortium** Recommendation. 2006. Disponível em: <http://www.w3pdf.com/W3cSpec/XML/2/REC-xml11-20060816.pdf>. Acesso em: 29 nov. 2019.

BREITMAN, K. K. **Web semântica*:* a internet do futuro**. Rio de Janeiro: LTC, 2005. 190 p.

BRILL, E. A simple rule-based part of speech tagger. In: **Proceedings of the Third Conference on Applied Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 152–155, 1992. Disponível em: <https://dl.acm.org/citation.cfm?doid=974499.974526>. Acesso em: 30 dez. 2019.

BRUYNE, P.; HERMAN, J.; SCHOUTHEETE, M. **Dinâmica da pesquisa em Ciências Sociais.** Rio de Janeiro: F. Alves, 1991

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science (1986-1998)**, v. 42, n. 5, p. 351, 1991. Disponível em: <https://search.proquest.com/docview/216897238?pq-origsite=gscholar>. Acesso em: 25 set. 2017.

CABALEIRO, B.; PEÑAS, A.; MANANDHAR, S. Grounding proposition stores for question answering over linked data. **Knowledge-Based Systems**, v. 128, p. 34-42, 2017.

CAMPOS, L. M.; CAMPOS, M. L. A. Aplicação de dados interligados abertos apoiada por ontologia. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO: ALÉM DAS NUVENS, EXPANDINDO AS FRONTEIRAS DA CIÊNCIA DA INFORMAÇÃO, 15.1: 3822-3841, Belo Horizonte, MG. **Anais eletrônicos...** Belo Horizonte, MG: ANCIB, 2014. Disponível em <http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8>. Acesso em: 16 mar. 2020

CAPUANO, E. A. O poder cognitivo das redes neurais artificiais modelo Art1 na recuperação da informação. **Ciência da Informação**, v. 38, n. 1, p. 9-30, 2009. Disponível em: <http://eprints.rclis.org/17764/1/Capuano-Art-v38n1-2009.pdf>. Acesso em: 26 dez. 2019.

CAROTHERS, G; SEABORNE, A. **RDF 1.1 N-Triples**. 2014. Disponível em: <https://www.w3.org/TR/n-triples/>. Acesso em: 03 jul. 2019

CASSETTARI, R. R. B. et al. COMPARAÇÃO DA LEI DE ZIPF EM CONTEÚDOS TEXTUAIS E DISCURSOS ORAIS. **El profesional de la información**, v. 24, n. 2, 2015. Disponível em: <https://www.researchgate.net/profile/Rosangela_Rodrigues5/publication/277930184_Comparacao_da_Lei_de_Zipf_em_conteudos_textuais_e_discursos_orais/links/5a731673aca2720bc0dac4ff/Comparacao-da-Lei-de-Zipf-em-conteudos-textuais-e-discursos-orais.pdf>. Acesso em: 18 jul. 2019.

CASTELLS, M. **A sociedade em rede**. São Paulo: Paz e Terra, 2007.

CASTELLS, M. **O poder da comunicação**. São Paulo: Paz e Terra, 2015.

CHARNIAK, E.; MCDERMOTT, D. **Introduction to artificial intelligence**. Pearson Education India, 1985.

CHIJINDU, E. V. C. Search in Artificial Intelligence Problem Solving. **African Journal of Computing & ICT**, v. 5, n. 5, p. 37, 2012. Disponível em: https://www.researchgate.net/publication/326039952_Search_In_Artificial_Intelligence_Problem_Solving. Acesso em: 30 dez. 2019.

CHOMSKY, Noam. **Syntactic structures**. Mouton Publisher: Berlin, 1957.

CODD, E. F. A relational model of data for large shared data banks. **Communications of the ACM**, v. 13, n. 6, p. 377-387, 1970.

CONEGLIAN, C. S. **MODELO COMPUTACIONAL DE RECUPERAÇÃO DA INFORMAÇÃO PARA REPOSITÓRIOS DIGITAIS UTILIZANDO ONTOLOGIAS**. Dissertação (Mestrado em Ciência da Informação). Faculdade de Filosofia e Ciências. Universidade Estadual Paulista. Marília. 2017. Disponível em: < https://repositorio.unesp.br/bitstream/handle/11449/148996/coneglian_cs_me_mar.pdf?sequence=3&isAllowed=y >. Acesso em: 13 mar. 2017.

CONEGLIAN, C. S. et al. A experiência do usuário nos mecanismos de busca Knowledge Graph e o Knowledge Vault. **Informação@ Profissões**, v. 6, n. 2, p. 35-59. Disponível em: <http://www.uel.br/revistas/uel/index.php/infoprof/issue/view/1469>. Acesso em: 16 nov. 2019.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Materialização da Web Semântica: um modelo de construção dinâmica de consultas baseados em mapeamento de ontologias. **Perspect. ciênc. inf**., Belo Horizonte , v. 23, n. 2, p. 33-49, June 2018 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362018000200033&lng=en&nrm=iso>. Acesso em: 04 jul. 2019. http://dx.doi.org/10.1590/1981-5344/2728.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das humanidades digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 48, p. 88-99, 2017.

COSTA, E.; CAMPELO, C.; SAMPAIO, L. Classificação automática de questões. Problema de matemática para aplicações do pensamento computacional na educação. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2018. p. 569.

CROCKFORD, D. **Introdução ao JSON**. 2019. Disponível em: <https://www.json.org/json-pt.html>. Acesso em: 03 jul. 2019.

CUNHA, I. M. R. F; KOBASHI, N. Y. Análise documentária e inteligência artificial. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 24, n. 1/4, p. 38-62, 1991.

DEVI, N. V.; PONNUSAMY, R. A Systematic Survey of Natural Language Processing (NLP) Approaches in Different Systems. **International Journal of Computer Sciences and Engineering**, v. 4, n. 7, p. 192-198, 2016. Disponível em: https://www.researchgate.net/profile/Ramalingam_Ponnusamy5/publication/322465983_A_Systematic_Survey_of_Natural_Language_Processing_NLP_Approaches_in_Different_Systems/links/5a59d8914585154502711b7a/A-Systematic-Survey-of-Natural-Language-Processing-NLP-Approaches-in-Different-Systems.pdf. Acesso em: 17 jul. 2019.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH. **Years Public Domain for the original Web Software** : message from Tim Berners-Lee. 2003. Disponível em: https://videos.cern.ch/record/2038525. Acesso em: 20 mar. 2020.

EUROPEANA. **Definition of the Europeana Data Model v5.2.6**. 2014. Disponível em: <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requireme

nts/EDM_Documentation/EDM%20Definition%20v5.2.6_01032015.pdf> Acesso em: 17 fev. 2016.

FACEBOOK. **Default Structured Search Queries on Online Social Networks**. US Pat.20130124542A1, 16 maio 2016. Disponível em: <https://patents.google.com/patent/US20130124542A1/en?q=facebook>;. Acesso em: 30 jun. 2019.

FERNEDA, E. **Introdução aos Modelos Computacionais de Recuperação de Informação**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2012.

FERNEDA, E. **Recuperação de Informação**: estudo sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003, 147 f. 2003. Tese (Doutorado em Ciências da Comunicação) - Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003.

FERNEDA, E.; DIAS, G. A. A Lógica Fuzzy aplicada à recuperação de informação. **Revista InterScientia**, v. 1, n. 1, p. 51-65, 2013. Disponível em: <https://periodicos.unipe.br/index.php/interscientia/article/view/24>. Acesso em: 26 dez. 2019.

FERRÁNDEZ, Oscar et al. Addressing ontology-based question answering with collections of user queries. **Information Processing & Management**, v. 45, n. 2, p. 175-188, 2009.

FERREIRA, J. A.; SANTOS, P. L. V. A. C. O modelo de dados resource description framework (RDF) e o seu papel na descrição de recursos. **Informação & Sociedade.** João Pessoa, v. 23, n. 2, p. 13-23, maio/ago. 2013. Disponível em: <http://www.periodicos.ufpb.br/ojs2/index.php/ies/article/view/15436> Acesso em: 17 fev. 2020.

FUSCO, E. **Modelos conceituais de dados como parte do processo da catalogação: perspectiva de uso dos FRBR no desenvolvimento de catálogos bibliográficos digitais.** 2010. 249f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

GANDON, F.; SCHREIBER, G. **RDF 1.1 XML Syntax**. 2014. Disponível em: <https://www.w3.org/TR/rdf-syntax-grammar/>. Acesso em: 03 jul. 2019.

GONZALEZ, M.; LIMA, V. LS. Recuperação de informação e processamento da linguagem natural. In: **XXIII Congresso da Sociedade Brasileira de Computaçao**. 2003. p. 347-395.

GOOGLE. **Entender como dados estruturados funcionam.** 2019. Disponível em: <https://developers.google.com/search/docs/guides/intro-structured-data?hl=pt-br>. Acesso em: 04 jul. 2019.

GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing**. Knowledge Systems Laboratory, Stanford University, 1993. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.6200 >. Acesso em:

GHAHRAMANI, Z. Probabilistic machine learning and artificial intelligence. **Nature**, v. 521, n. 7553, p. 452-459, 2015. Disponível em:

<https://www.repository.cam.ac.uk/bitstream/handle/1810/248538/Ghahramani%25202015%2520Nature.pdf?sequence=1>. Acesso em: 31 dez. 2019.

GUARINO, N. Formal ontology in information systems. In: **Proceedings of the first international conference** (FOIS'98), June 6-8, Trento, Italy. IOS press, 1998.

GUIMARÃES, H. R. **RECUPERAÇÃO DE INFORMAÇÕES MUSICAIS**: UMA ABORDAGEM UTILIZANDO DEEP LEARNING. 2018. Monografia. Universidade Federal do Rio de Janeiro. Disponível em: <http://monografias.poli.ufrj.br/monografias/monopoli10025687.pdf>. Acesso em: 26 dez. 2019.

HAUGELAND, J. **Artificial intelligence**: The very idea. MIT press, 1985.

HEATH, T.; BIZER, C. **Linked Data**: Evolving the Web into a Global Data Space (1st edition). EUA: Morgan & Claypool, 2011.

HINTON, G.; VINYALS, O.; DEAN, J.. Distilling the knowledge in a neural network. **arXiv** preprint arXiv:1503.02531, 2015. Disponivel em: <https://arxiv.org/pdf/1503.02531.pdf>. Acesso em: 30 dez. 2019.

HJORLAND, B. The foundation of the concept of relevance. **Journal of the American Society for Information Science and Technology**, p. 217-237, v. 61, 2009. Disponível em: <http://onlinelibrary.wiley.com/wol1/doi/10.1002/asi.21261/full>. Acesso em: 09 fev. 2020.

HORROCKS, I. et al. **SWRL**: A Semantic Web Rule Language Combining OWL and RuleML. 2004. Disponível em: <https://www.w3.org/Submission/SWRL/>. Acesso em: 04 jul. 2019.

IBM. **Build apps with natural language processing**. Disponível em: <https://www.ibm.com/watson/natural-language-processing>. Acesso em: 08 jul. 2019.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. Novatec Editora, 2015.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015. Disponível em: <https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf>. Acesso em: 31 dez. 2019.

JOSHI, A. K. Natural language processing. **Science**, v. 253, n. 5025, p. 1242-1249, 1991. Disponível em: <https://go.galegroup.com/ps/i.do?id=GALE%7CA11360629&sid=googleScholar&v=2.1&it=r&linkaccess=fulltext&issn=00368075&p=AONE&sw=w&casa_token=mhYeTDNOle8AAAAA:UvJ5x_ARFpiOzKkciEkXdootzxDDfccN03ve-_y7iftcQSz3A6Tj2gREmJEuYA2awaj5rzSp9rQ>. Acesso em: 05 jul. 2019.

KEPUSKA, V.; BOHOUTA, G.. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In: **2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)**. IEEE, 2018. p. 99-103.

KHURANA, Diksha et al. Natural language processing: State of the art, current trends and challenges. **arXiv preprint arXiv:1708.05148**, 2017. Disponível em: <https://arxiv.org/ftp/arxiv/papers/1708/1708.05148.pdf>. Acesso em: 20 maio 2020.

KLYNE, G.; CARROLL, J. **Resource Description Framework (RDF)**: Concepts and Abstract Syntax. Disponível em: <https://www.w3.org/TR/rdf-concepts/>. Acesso em: 15 maio 2020.

KUMAR, G. S.; ZAYARAZ, G. Concept relation extraction using Naïve Bayes classifier for ontology-based *Question Answering* systems. **Journal of King Saud University-Computer and Information Sciences**, v. 27, n. 1, p. 13-24, 2015. Disponível em: <http://ac.els-cdn.com/S1319157814000020/1-s2.0-S1319157814000020-main.pdf?_tid=897d5a8a-6701-11e6- 944e-00000aab0f26&acdnat=1471716878_680a3eddb48f7f42cc5ef92ed05ac125>. Acesso em: 17 ago. 2019.

KURZWEIL, R. **The age of intelligent machines**. Cambridge, MA: MIT press, 1990. Disponível em: <https://mitpress.mit.edu/books/age-intelligent-machines>. Acesso em: 14 mar. 2020.

LANCASTER, F. W.; WARNER, A. J. **Information retrieval today**. Arlington: Information Resources Press, 1993.

LAMPROPOULOS, A. S.; TSIHRINTZIS, G. A. **Machine learning paradigms**. Applications in recommender systems. Switzerland: Springer International Publishing, 2015. Disponível em: <https://link.springer.com/book/10.1007%2F978-3-319-19135-5>. Acesso em: 31 dez. 2019.

LASSILA, O.; MCGUINESS, D. **The Role of Frame-Based Representation on the Semantic Web**. Disponível em: <http://www-ksl.stanford.edu/pub/KSL_Reports/KSL-01-02.html>. Acesso em: 30 dez. 2019.

LATHAM, A.; CROCKETT, K.; MCLEAN, D. An adaptation algorithm for an intelligent natural language tutoring system. **Computers & Education**, 71, 97–110. 2014 doi: 10.1016/j.compedu.2013.09.014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360131513002698?via%3Dihub>. Acesso em: 22 abr. 2020.

LI, D.; DU, Y. **Artificial intelligence with uncertainty**. CRC press, 2017. Disponível em: <https://www.crcpress.com/Artificial-Intelligence-with-Uncertainty/Li-Du/p/book/9781498776264>. Acesso em: 14 mar. 2020.

LIDDY, E. D. Natural language processing. In **Encyclopedia of Library and Information Science**, 2nd Ed. NY. Marcel Decker, Inc. 2001. Disponível em: <https://surface.syr.edu/cgi/viewcontent.cgi?referer=http://scholar.google.com.br/&httpsredir=1&article=1019&context=cnlp>. Acesso em: 05 jul. 2019.

LOPEZ, V. et al. Evaluating *Question Answering* over *Linked Data*. Web Semantics: Science, Services and Agents on the World Wide Web, v. 21, p. 3-13, 2013. Disponível em: < https://www.sciencedirect.com/science/article/abs/pii/S157082681300022X>. Acesso em: 14 mar. 2020.

LUGER, G. F.; STUBBLEFIELD, W. A. **Artificial intelligence**: structures and strategies for complex problem solving. Benjamin/Cummings, Redwood City, California, 1993.

LUZ, F. F. **Consulta a Ontologias utilizando Linguagem Natural Controlada**. 2013. Dissertação (Mestrado em Ciência da Computação). Instituto de Matemática e Estatística. Universidade de São Paulo. São Paulo. 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-02012014-224412/publico/fabianoDissertacao.pdf>. Disponível em: 05 jul. 2019.

MAIMONE, G. D. et al. Reflexões acerca das relações entre representação temática e descritiva. **Informação & Sociedade**: Estudos, João Pessoa, v. 21, n. 1, p. 27-35, 2011. Disponível em: http://www.brapci.inf.br/index.php/article/view/0000010197/8769e35e967e42828981ae72f80dacf8/. Acesso em: 27 dez. 2019.

MENDEZ, E. GREENBERG, J. *Linked Data* for open vocabularies and HIVE's global framework. **El profesional de la información**, 2012, mayo-junio, v. 21, n. 3. Disponível em: <http://recyt.fecyt.es/index.php/EPI/article/view/epi.2012.may.03/17916>

MICHAELIS. **Moderno Dicionário da Língua Portuguesa**. Disponível em: <http://michaelis.uol.com.br/>. Acesso em: 15 jul. 2019.

MILES, A.; BECHHOFER, S. **SKOS simple knowledge organization system reference**. 2009. Disponível em: <https://www.w3.org/TR/skos-reference/>. Acesso em: 04 jul. 2019.

MONTEIRO, S. D. Knowledge Graph e a significação: novos agenciamentos semióticos dos índices contemporâneos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15, João Pessoa, PB. 2015.**Anais eletrônicos**... João Pessoa, PB: ANCIB, 2015. Disponível em:<http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/3025> Acesso em: 05 abr. 2018.

MONTEIRO, S. D. et al. Sistemas de recuperação da informação e o conceito de relevância nos mecanismos de busca: semântica e significação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 50, p. 161-175, 2017. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/viewFile/1518-2924.2017v22n50p161/34700>. Acesso em: 18 jul. 2019.

MOOERS, C.N. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, 2, 2 32. 1951. Disponível em: < https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090020107>. Acesso em: 14 mar. 2020.

NOVELLINO, M. S. F. Instrumentos e metodologias de representação da informação. **Informação & Informação**, v. 1, n. 2, p. 37-45, 1996. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/download/1603/1358. Acesso em: 26 dez. 2019.

NUNES, M. G. V. **Processamento de línguas naturais: para quê e para quem?**. ICMC-USP, 2008. Disponível: <http://wiki.icmc.usp.br/images/5/55/ND_73.pdf>. Acesso em> 08 jul. 2019.

OLIVEIRA, H. P. C.; VIDOTTI, S. A. B. G. Método Quadripolar: Aplicação em Pesquisas Informacionais e Tecnológicas. In: PINTO, V. B.; VIDOTTI, S. A. B. G.; CAVALCANTE, L. E. (Org.). **Aplicabilidades metodológicas em ciência da informação.** Fortaleza: Edições UFC, 2015. cap. 2, p. 35-47.

RAMALHO, R. A. S. **Web semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da ciência da informação**. 2006. 120 f. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2006. Disponível em: <http://hdl.handle.net/11449/93709>. Acesso em

RAMALHO, R. A. S.; VIDOTTI, S. A. B. G.; FUJITA, M. S. L. Web semântica: uma investigação sob o olhar da Ciência da Informação. **DataGramaZero–Rev**. v. 8, n. 6, 2007. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/_repositorio/2010/01/pdf_7557383cd1_0007573.pdf>. Acesso em: 25 set. 2017.

REIS, H. M.; JAQUES, P. A.; ISOTANI, S. Sistemas tutores inteligentes que detectam as emoções dos estudantes: um mapeamento sistemático. **Brazilian Journal of Computers in Education**, v. 26, n. 03, p. 76, 2018. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/7184>. Acesso em: 27 dez. 2019.

RIBEIRO, A. C.; FRAZÃO, R.; SA, J. O. Machine Learning Puzzles: How to select Use Cases, Algorithms and Technologies? **Association for Information Systems. AIS Electronic Library (AISeL)**. 2018. Disponível em: <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1029&context=capsi2018>. Acesso em: 31 dez. 2019.

RICH E.; KNIGHT K. **Intelligence Artificial**. 1991. McGraw-Hill, New York. 1992.

ROA-MARTINEZ, S. M. **Da information findability à image findability**: aportes da polirrepresentação, recuperação e comportamento de busca. 2019. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence**: a modern approach. Malaysia; Pearson Education Limited, 2016.

SANTAREM SEGUNDO, J. E. **Representação Iterativa: um modelo para Repositórios Digitais**. 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010. Disponível em: <https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/santaremsegundo_je_do_mar.pdf>. Acesso em: 04 jul. 2019.

SANTAREM SEGUNDO, J. E. Tim Berners-Lee e a Ciência da Informação: do hipertexto à web semântica. In: **Os pensadores e a Ciência da Informação.** Rio de Janeiro: E-papers, 2012, v.1, p. 101-110. Disponível em: < https://repositorio.usp.br/item/002305024>. Acesso em: 14 mar. 2020.

SANTARÉM SEGUNDO, J. E. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas. **Tendências da Pesquisa Brasileira em Ciência da**

**Informação**, v. 8, n. 2, p. 219-239, 2015. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207>. Acesso em: 05 jul. 2019.

SANTARÉM SEGUNDO, J. E. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 8, n. 2, p.219-239, 2015. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207>. Acesso em: 21 fev. 2016.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução à recuperação de dados usando Sparql. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15, Belo Horizonte, MG, 2014. **Anais eletrônicos...** Belo Horizonte, MG: ANCIB, 2014. Disponível em <http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt8> Acesso em: 2 fev. 2019.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Tecnologias da Web Semântica aplicadas a organização do conhecimento: padrão SKOS para construção e uso de vocabulários controlados descentralizados. In: José Augusto Chaves Guimarães; Vera Dodebei. (Org.). **Organização do Conhecimento e Diversidade Cultural**. 1.ed. Marília: Fundepe, v. 3, p. 224-233, 2015. Disponível em: <http://isko-brasil.org.br/wp-content/uploads/2015/09/Organiza%C3%A7%C3%A3o-do-Conhecimento-e-Diversidade-Cultural-ISKO-BRASIL-2015.pdf>. Acesso em: 25 set. 2017.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação & Informação**, v. 21, n. 2, p. 217-244, 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/viewFile/26417/20131>. Acesso em: 23 mar. 2018.

SANTAREM SEGUNDO, J. E.; SOUZA, J.; CONEGLIAN, C. S. Web Semântica: introdução a recursos de visualização de dados em formatos gráficos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15, João Pessoa, PB. 2015. **Anais eletrônicos...** João Pessoa, PB: ANCIB, 2015. Disponível em: <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/2780> Acesso em: 28 maio 2020.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**. vol. 24, n. 1, 1995. Disponível em: <http://www.brapci.ufpr.br/brapci/_repositorio/2010/03/pdf_dd085d2c4b_0008887.pdf>. Acesso em: 25 set. 2017.

SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science. **Journal of the American Society for information science**, v. 26, n. 6, p. 321-343, 1975. Disponível em: https://s3.amazonaws.com/academia.edu.documents/30771284/relevanceSaracevic.pdf?response-content-disposition=inline%3B%20filename%3DRelevance_A_review_of_and_a_framework_fo.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20190718%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190718T214830Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-

Signature=c38f6e85c35252a6bd455f979f9e3470770944d675b88b70d4ea08d4573eca89. Acesso em: 18 jul. 2019.

SCHALKOFF, R. J. **Artificial intelligence**: an engineering approach. New York: McGraw-Hill, 1990.

SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. The semantic web revisited. **IEEE intelligent systems**, v. 21, n. 3, p. 96-101, 2006. Disponível em: < https://eprints.soton.ac.uk/262614/1/Semantic_Web_Revisted.pdf>. Acesso em: 14 mar. 2020.

SILVA, W. B.; CALUMBY, R. T.. Recuperação de Imagens na Web com Fusão Adaptativa de Credibilidade Baseada em Algoritmos Genéticos. In: **Anais Estendidos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web**. SBC, 2018. p. 53-56. Disponível em: https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/4056/3996. Acesso em: 26 dez. 2019.

SILVA, A. M. **A informação**: da compreensão do fenómeno e construção do objecto científico. Porto: Ed. Afrontamento, 2006.

SILVA, A. M. O método quadripolar e a pesquisa em ciência da informação. **Prisma. com**, n. 26, 2017. Disponível em: <https://pentaho.letras.up.pt/ojs/index.php/prismacom/article/viewFile/1861/1694>. Acesso em: 25 set. 2017.

SILVA, A. M. RIBEIRO, F. **Das "Ciências documentais à Ciência da Informação**: ensaio epistemológico para um novo modelo curricular. Porto: Ed. Afrontamento, 2002.

SILVA, B. C. D. et al. **Introdução ao processamento das línguas naturais e Algumas Aplicações**. NILC-USP: São Carlo. 2007. Disponível em: <http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0710-DiasDaSilvaEtAl.pdf>. Acesso em: 05 jul. 2019.

SILVA, R. E.; SANTOS, P. L. V. A. da C.; FERNEDA, E. Modelos de recuperação de informação e web semântica: a questão da relevância. **Informação & Informação**, v. 18, n. 3, p. 27-44, 2013. Disponível em: <http://hdl.handle.net/11449/114705>. Acesso em: 19 jul. 2019.

SINGHAL, A. **Introducing the knowledge graph**: things, not strings. 2012. Disponível em: <http://googleblog.blogspot.com.br/2012/05/introducingknowledge-graph-things-not.html>. Acesso em: 05 abr. 2018.

SOUZA, R. F. Organização e representação de áreas do conhecimento em ciência e tecnologia: princípios de agregação em grandes áreas segundo diferentes contextos de produção e uso de informação 10.5007/1518-2924.2006 v11nesp1p27. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, v. 11, n. 1, p. 27-41, 2006. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/337. Disponível em: 26 dez. 2019.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação,** v. 33, n. 1, p. 132-141, 2004. Disponível em: <http://revista.ibict.br/ciinf/article/view/1077/1176>. Acesso em: 25 set. 2017.

SOUZA, R.R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161-173, maio/ago. 2006. Disponível em: http://www.scielo.br/pdf/%0D/pci/v11n2/v11n2a02.pdf. Acesso em: 18 jul. 2019.

SPORNY, M. et al. **JSON-LD 1.1**. A JSON-based Serialization for *Linked Data*. 2019. Disponível em: <https://json-ld.org/spec/latest/json-ld/>. Acesso em: 03 jul. 2019.

THE LINKED OPEN DATA CLOUD. **LOD Cloud**. 2019. Disponível em: <https://lod-cloud.net/versions/2019-03-29/lod-cloud.png>. Acesso em: 05 jul. 2019.

UNGER, C. et al. Template-based *Question Answering* over RDF data. In: **Proceedings of the 21st international conference on World Wide Web**. ACM, 2012. p. 639-648. Disponível em: https://pub.uni-bielefeld.de/download/2495397/2526223/template-based_question.pdf. Acesso em: 14 mar. 2020.

VECHIATO, F. L. **Encontrabilidade da informação**: contributo para uma conceituação no campo da ciência da informação. 2013. 206 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2013. Disponível em: <https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/Tese_de_Doutorado_-_Fernando_Luiz_Vechiato.pdf>. Acesso em: 25 set. 2017.

VECHIATO, F. L; VIDOTTI, S. A. B. G. **Encontrabilidade da informação**. Coleção PROPG Digital (UNESP), 2014. Disponível em: < https://repositorio.unesp.br/bitstream/handle/11449/126218/ISBN9788579835865.pdf?sequence=1>. Acesso em: 20 maio 2020.

VIEIRA, R; LOPES, L. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. **EM CORPORA**, p. 183, 2010. Disponível: <https://s3.amazonaws.com/academia.edu.documents/33579040/linguagensespecializadasemcorpora.pdf?response-content-disposition=inline%3B%20filename%3DLINGUAGENS_ESPECIALIZADAS.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20190705%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20190705T174023Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=3b8d5c040735319be304f7cd4b8c23a3a6ee1f42715e40a4781c08c07cc6b7ef#page=184>. Acesso em: 19 dez. 2019.

VIOTTI, E. C. **Introdução aos estudos lingüísticos.** Florianópolis: Universidade Federal de Santa Catarina. 2008. Disponível em: http://www.libras.ufsc.br/colecaoLetrasLibras/eixoFormacaoBasica/estudosLinguisticos/assets/317/TEXTO_BASE_-_VERSAO_REVISADA.pdf. Acesso em: 02 jan. 2020.

WALTER, S. et al. Evaluation of a layered approach to Question Answering over Linked Data. In**: The Semantic Web–ISWC** 2012. Springer Berlin Heidelberg, 2012. p. 362-374. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-35173-0_25. Acesso em: 14 mar. 2020.

WINSTON, P. H. **Artificial Intelligence Addison-Wesley**. Readig, MA, 1992.

WORLD WIDE CONSORTIUM. OWL 2 **Web Ontology Language Document Overview (Second Edition).** 2012b. Disponível em: <https://www.w3.org/TR/owl2-overview/>. Acesso em: 04 jul. 2019.

WORLD WIDE WEB CONSORTIUM. **Extensible Markup Language (XML)**. 2011. Disponível em: <https://www.w3.org/XML//>. Acesso em: 03 jul. 2019.

WORLD WIDE WEB CONSORTIUM. **OWL Web Ontology Language**. 2012a. Disponível em: <https://www.w3.org/OWL/>.

WORLD WIDE WEB CONSORTIUM. **OWL**: Web Ontology Language (OWL). 2012. Disponível em: <https://www.w3.org/OWL/>. Acesso em: 25 set. 2017.

WORLD WIDE WEB CONSORTIUM. **RDF**. Resource Description Framework (RDF). 2014. Disponível em: <https://www.w3.org/RDF/>. Acesso em: 29 maio 2019.

WORLD WIDE WEB CONSORTIUM. **SKOS Simple Knowledge Organization System - Home Page**. 2012c. Disponível em: <https://www.w3.org/2004/02/skos/>. Acesso em: 07 mar. 2020.

WORLD WIDE WEB CONSORTIUM. **W3C Opens Data on the Web with SPARQL**. 2007. Disponível em: <https://www.w3.org/2007/12/sparql-pressrelease.html.en>. Acesso em: 04 jul. 2019.

XIE, X. et al. Research and implementation of automatic question answering system based on ontology. In: The 27th Chinese Control and Decision Conference (2015 CCDC). **IEEE**, 2015. p. 1366-1370.

ZIPF, G. K. **Human behavior and the principle of least effort: An introduction to human ecology**. Ravenio Books, 2016.