

UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Faculdade de Filosofia e Ciências
Campus Marília

CAIO SARAIVA CONEGLIAN

**MODELO COMPUTACIONAL DE RECUPERAÇÃO DA INFORMAÇÃO
PARA REPOSITÓRIOS DIGITAIS UTILIZANDO ONTOLOGIAS**

Marília - SP

2017

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

CAIO SARAIVA CONEGLIAN

**MODELO COMPUTACIONAL DE RECUPERAÇÃO DA INFORMAÇÃO
PARA REPOSITÓRIOS DIGITAIS UTILIZANDO ONTOLOGIAS**

Dissertação 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 mestre em Ciência da Informação.

Área de Concentração: Informação, Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Financiamento: CAPES e FAPESP (processo nº 2015/01517-2)

Orientador: Dr. José Eduardo Santarem Segundo

Marília – SP

2017

Coneglian, Caio Saraiva.

C747m Modelo computacional de recuperação da informação para repositórios digitais utilizando ontologias / Caio Saraiva Coneglian. – Marília, 2017.

143 f. ; 30 cm.

Orientador: José Eduardo Santarém Segundo.

Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, 2017.

Bibliografia: f. 136-143.

1. Web semântica. 2. Ontologias (Recuperação da informação). 3. Repositórios institucionais. 4. Tecnologia da informação. I. Título.

CDD 004.6

CAIO SARAIVA CONEGLIAN

MODELO COMPUTACIONAL DE RECUPERAÇÃO DA INFORMAÇÃO PARA
REPOSITÓRIOS DIGITAIS UTILIZANDO ONTOLOGIAS

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

Área de Concentração: Informação,
Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Data da defesa: 17 de fevereiro de 2017.

BANCA EXAMINADORA

José Eduardo Santarem Segundo (Orientador)
Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/FFC

Silvana Aparecida Borsetti Gregorio Vidotti
Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/FFC

Elvis Fusco
Docente do Departamento de Ciência da Computação da UNIVEM/Marília

Suplentes

Ricardo Cesar Gonçalves Santana
Docente do Programa de Pós-Graduação em Ciência da Informação da UNESP/FFC

Leonardo Castro Botega
Docente do Departamento de Ciência da Computação da UNIVEM/Marília

AGRADECIMENTOS

Agradeço à Natalia, que muito me auxiliou, sendo uma grande parceira e companheira, essencial na minha vida.

Aos meus pais, Ana Maria e José Artur, pelo apoio e pela compreensão, que direta e indiretamente foram essenciais para a realização desse momento. Ao meu irmão, Fernando, a minha sobrinha, Maria Fernanda, e a todos os meus familiares.

Ao Anderson, Felipe, Marcelo, Mariana, Maycon e Silvio, verdadeiros amigos, com quem dividi diversos bons momentos.

Ao meu orientador José Eduardo Santarém Segundo, pela orientação e pela ajuda no desenvolvimento da pesquisa, fundamental para eu conseguir aprimorar um pouco minha vivência científica.

À Silvana Vidotti, pelas contribuições para a pesquisa e por participar da minha banca de defesa, além de todos os ensinamentos, fundamentais para eu compreender melhor a Ciência da Informação e realizar o mestrado.

Ao Elvis Fusco, pelas contribuições para a pesquisa e por participar da minha banca de defesa, além da ajuda e dos ensinamentos para delimitar meu projeto e oportunidades me concedidas durante o período.

A todos os docentes do Programa de Pós-Graduação de Ciência da Informação da UNESP Marília, por todos os ensinamentos me passados.

À Maria Lúcia Balestrieri, pela correção ortográfica.

Aos meus amigos e colegas do PPGCI, em especial a Ana Maria, Paula, Sandra, Felipe, Cecílio, Edgar, Jessica, Larissa e Victor, que muito me ajudaram, tanto academicamente, quanto pessoalmente, a passar por este período.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (Número do processo: 2015/01517-2) e a CAPES pelo apoio ao trabalho.

À Deus, por tudo.

RESUMO

A evolução das Tecnologias da Informação e Comunicação causaram um aumento exponencial na produção e disseminação de dados na Internet. Dentre estas informações, inclui-se a produção científica que vive hoje um momento de transição, em que os documentos deixaram de ser apenas publicados em livros e revistas impressas e passaram a se espalhar pela rede. A partir disto, surgiu uma nova tecnologia chamada de repositórios digitais, em que são armazenados documentos em diversos formatos. Junto com o surgimento dos repositórios digitais, cresceu o desafio da recuperação destes documentos de maneira eficiente, ou seja, como a máquina poderá compreender o que o usuário procura, para fornecer os documentos que este usuário necessita. Neste âmbito, a Web Semântica surgiu visando possibilitar com que os computadores consigam compreender o contexto em que as informações criadas pelos usuários se encontram, tendo em suas ferramentas a base para tornar tal propósito real. No contexto dos repositórios digitais, esta pesquisa tem como objetivo aprimorar o processo de recuperação de informação nesses ambientes informacionais por meio da utilização do conceito de representações semânticas no uso de ontologias de domínio, que permita uma maior aderência na intersecção entre os itens bibliográficos e as necessidades informacionais dos usuários. Para atingir tais objetivos, utilizou-se uma metodologia de natureza quantitativa, em que se criou um modelo utilizando conceitos e tecnologias da Web Semântica para contextualizar o domínio da busca realizada pelo usuário. Como resultados, verificou-se que as relações das ontologias podem ser extraídas com eficiência por meio de um motor de geração de consultas SPARQL, que consegue localizar um termo na ontologia, bem como extrair as relações desse termo. Além disso, foram traçadas as ações que as propriedades do OWL devem possuir, no âmbito da recuperação da informação, para que assim possam ser identificadas com maior precisão as relações que um termo de busca possui frente a uma ontologia, permitindo a geração de uma nova expressão de busca, contendo um maior número de argumentos. Outro resultado obtido, diz respeito a interoperabilidade em repositórios digitais, que possibilitou identificar a integração e a recuperação dos metadados dos documentos dos repositórios digitais e a ferramenta tratando das questões semânticas. O trabalho propôs a interatividade na escolha das fontes informacionais, em que o usuário escolhe os repositórios em que seria realizada a busca, bem como cadastrar um repositório, caso este não tenha sido utilizado anteriormente. Conclui-se que a inserção de semântica em processos de recuperação de informação pode ocorrer por meio do modelo proposto, que se baseia essencialmente nas tecnologias e nos conceitos da Web Semântica, especialmente as ontologias, como um artefato capaz de explicitar o contexto em que os termos se encontram.

Palavras-chave: web semântica; ontologias; repositórios digitais; recuperação de informação.

ABSTRACT

The evolution of Information and Communication Technologies has caused an exponential increase in the production and dissemination of data on the Internet. Among these data, we include the scientific production that is now in a moment of transition, in which the documents are no longer only published in printed books and magazines, and are now spread throughout the network. From this, a new technology has emerged called digital repositories, in which documents are stored in various formats. Along with the emergence of digital repositories, the challenge of recovering these documents has grown in an efficient way, i.e., how the machine can understand what the user is looking for, to provide the documents that this user needs. In this context, the Semantic Web came about in order to enable computers to understand the context in which the information created by the users meet, having in their tools the basis to make such a real purpose. In the context of Digital Repositories, this research aims to improve the Information Retrieval process in these informational environments through the use of the concept of semantic representations in the use of domain ontologies, which allows greater adherence at the intersection between bibliographic items and Information needs of users. In order to achieve these objectives, a quantitative methodology was used, in which a digital repositories interoperability model was created, using Semantic Web concepts and technologies to contextualize the search domain performed by the user. As results, it was verified that the relationships of the ontologies can be extracted efficiently by means of a SPARQL query engine, that is able to locate a term in the ontology, as well as to extract the relations of this term. In addition, we have outlined the actions that OWL properties must possess, in the context of Information Retrieval, so that the relationships that a search term has against an ontology can be identified more precisely, allowing the generation of a new Search expression, containing a greater number of arguments. Another result obtained concerns interoperability in digital repositories, which made it possible to identify the integration of the retrieval of the metadata of the digital repositories documents and the tool dealing with semantic issues. The work proposed the interactivity in the choice of informational sources, in which the user could choose the repositories in which the search would be carried out, as well as register a repository, if it had not previously been used. It is concluded that the insertion of semantics in Information Retrieval processes can occur through the proposed model, which is based essentially on the technologies and concepts of the Semantic Web, especially the ontologies, as an artifact capable of explaining the context in which the terms occur.

Keywords: semantic web; ontologies; digital repositories; information retrieval.

LISTA DE ILUSTRAÇÕES

Figura 1 - Arquitetura Padrão do Protocolo OAI-MH	30
Figura 2 - Camadas da Web Semântica.....	37
Figura 3 – Exemplo - representação gráfica do RDF	43
Figura 4 - Representação RDF/XML	43
Figura 5 - Exemplo de uma consulta SPARQL.....	59
Figura 6 - Funcionamento das cláusulas <i>where</i> e <i>select</i>	60
Figura 7 - Modelo Semântico de Interoperabilidade entre Repositórios.....	72
Figura 8 - Passos para a construção da expressão de busca automática.....	76
Figura 9 - Exemplo de grafo com nó em branco	79
Figura 10 - Diagrama de atividades do modelo.....	96
Figura 11 - Tela de pesquisa.....	98
Figura 12 - Tela com as relações encontradas	99
Figura 13 - Tela de resultados	100
Figura 14 - Tela dos metadados do artigo	100
Figura 15 - Tela de cadastro do repositório	101
Figura 16 - Tela de Login	102
Figura 17 - Relações escolhidas entre o termo de busca e a ontologia	105
Figura 18 - Lógica de inserção dos termos na expressão de busca	105
Figura 19 - Exemplo de expressão de busca utilizando o termo <i>Linked Data</i>	106
Figura 20 - Função SPARQL de localizar termo	108
Figura 21 - Função SPARQL localizar relações	108
Figura 22 - Consulta SPARQL com nós em branco.....	109
Figura 23 - Múltiplos nós brancos dentro de uma ontologia representada em grafo ...	110
Figura 24 - Consulta de simetria gerada.....	114
Figura 25 - Fragmento das relações existentes em uma consulta.....	115
Figura 26 - Fragmento do registro retornado pelo provedor de dados	117
Figura 27 - Fragmento de um registro XML	117
Figura 28 - Síntese do processo de coleta e de armazenamento dos metadados	119
Figura 29 - Interface de busca Apache Solr	120
Figura 30 - Funcionamento do módulo de indexação e de busca.....	122
Figura 31 - Definição das configurações da demonstração do protótipo	123

Figura 32 - Relações entre o termo “Weapon System” e a ontologia definida	124
Figura 33 - Nova expressão de busca gerada	124
Figura 34 - Listagem dos resultados obtidos	125
Figura 35 - Registro completo de um objeto do repositório.....	126

LISTA DE QUADROS

Quadro 1 - Elementos do Dublin Core	26
Quadro 2 - Principais classes RDF Schema	44
Quadro 3 - Propriedades RDF Schema.....	45
Quadro 4 - Propriedades Especiais OWL.....	54
Quadro 5 - Exemplo de resultado da consulta apresentada na Figura 6.....	61
Quadro 6 - Comportamento para recuperação da informação para propriedades de classes da OWL	82
Quadro 7 - Ações das propriedades de classes	85
Quadro 8 - Comportamento para recuperação da informação para propriedades de propriedade do OWL	87
Quadro 9 - Ações das propriedades de propriedades	90
Quadro 10 - Consulta SPARQL genérica para cada divisão de propriedade	111

LISTA DE ABREVIATURAS E SIGLAS

American National Standards Institute / The National Information Standards Organization (ANSI/NISO)

Dublin Core Metadata Initiative (DCMI)

eXtensible Markup Language (XML)

Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

International Organization for Standardization (ISO)

Internet Engineering Task Force Request for Comments (IETF RFC)

JavaServer Faces (JSF)

Machine Readable Cataloging (MARC)

Open Access Initiative (OAI)

Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)

Platform for Internet Content Selection (PICS)

Registry of Open Access Repositories (ROAR)

Resource Description Framework (RDF)

Uniform Resource Identifier (URI)

Uniform Resource Locator (URL)

Uniform Resource Name (URN)

Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)

Universidade Federal do Rio Grande do Sul (UFRGS)

Web Ontology Language (OWL)

World Wide Web Consortium (W3C)

SUMÁRIO

1 INTRODUÇÃO.....	12
1.1 Problema.....	14
1.2 Objetivos.....	15
1.3 Procedimentos Metodológicos.....	15
1.4 Justificativa.....	16
1.5 Trabalhos Correlatos.....	17
2 REPOSITÓRIOS DIGITAIS.....	20
2.1 Metadados.....	23
2.2 Protocolo OAI-PMH.....	28
3 WEB SEMÂNTICA.....	32
3.1 Conceitos e Tecnologias da Web Semântica.....	35
3.1.1 URI.....	38
3.1.2 XML.....	39
3.1.3 RDF e RDF Schema.....	40
3.1.4 Ontologias.....	46
3.1.5 SPARQL.....	57
4 RECUPERAÇÃO DA INFORMAÇÃO.....	62
4.1 Modelos de Recuperação de Informação.....	64
4.2 Apache SOLR.....	66
5 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO PARA REPOSITÓRIOS BASEADO EM ONTOLOGIAS.....	70
5.1 Camada de Apresentação.....	74
5.2 Camada de Controle.....	75
5.3 Camada Semântica.....	76
5.3.1 OWL na Recuperação de Informação.....	80
5.4 Camada de Dados.....	91
6 IMPLEMENTAÇÃO DO PROTÓTIPO.....	94
6.1 Camada de Apresentação.....	97
6.2 Camada de Controle.....	102
6.3 Camada Semântica.....	106
6.4 Camada de Dados.....	115
6.4.1 Coleta e Armazenamento dos Metadados.....	116
6.4.2 Processo de Recuperação da Informação no Provedor de Serviços.....	120

6.5 Demonstração de funcionamento.....	122
7 TRABALHOS FUTUROS	127
8 CONSIDERAÇÕES FINAIS	129
REFERÊNCIAS	136

1 INTRODUÇÃO

As Tecnologias de Informação e Comunicação estão mudando o modo como as pessoas interagem com a realidade. Há uma série de equipamentos e de sistemas informacionais trazendo facilidades nos mais diversos contextos da sociedade. Uma consequência deste processo concerne à disseminação e à mobilidade das informações, em que as variadas áreas do conhecimento, como as Engenharias, a Ciência da Computação, a Saúde e a Filosofia estão facilitando o acesso a suas pesquisas e a seus conhecimentos.

Nessa seara, a Ciência da Informação tem uma importância ímpar, ao apresentar uma interface de comunicação com as demais disciplinas, tendo uma participação efetiva nas transformações vivenciadas (SANTAREM SEGUNDO, 2010). Tais questões podem ser identificadas no texto de Saracevic (1995), em que o autor discorre a respeito da interdisciplinaridade da Ciência da Informação, relacionando a Biblioteconomia, a Ciência da Computação, a Ciência Cognitiva e a Comunicação.

Por meio das reflexões expostas por Saracevic, nota-se que a Ciência da Informação pode trazer contribuições essenciais para a maneira como a informação na conjuntura atual deve ser tratada, organizada e utilizada. Essas características podem ser identificadas em uma das principais definições dada por Borko sobre a Ciência da Informação, apontando que “[...] é a disciplina que investiga as propriedades e o comportamento da informação, as forças que governam os fluxos informacionais, e os significados do processamento da informação, para a acessibilidade e a usabilidade otimizadas” (BORKO, 1968, p.1, tradução nossa).

No âmbito das transformações vivenciadas pela sociedade, os efeitos provocados pela Internet têm tido a maior notoriedade. Um dos motivos para tal fato é que este ambiente é capaz de realizar a comunicação entre indivíduos, organizações e tecnologia de uma maneira antes inalcançável.

Um exemplo do efeito que a Internet provocou na disseminação da informação diz respeito às alterações nos suportes em que as comunicações científicas passaram a ser difundidas; assim, a publicação e a disponibilização das informações deixaram de ser exclusivamente em suporte de papel, passando a ocorrer em ambientes digitais, como portais de revistas científicas e repositórios digitais de acesso aberto.

As facilidades que os meios digitais estão trazendo para a difusão da informação contrastam com uma série de obstáculos que passaram a existir, em destaque o aumento

exponencial na geração de informação, que provoca a necessidade do desenvolvimento de novas tecnologias de recuperação da informação, visando localizar os documentos que atendam às necessidades informacionais dos usuários.

Um ferramental criado para tratar tais questões são os repositórios digitais, definidos como sistemas de informação que armazenam arquivos digitais para futura recuperação, incorporando a facilidade de comunicação, de colaboração e de interações dinâmicas entre o usuário e um vasto universo (SAYÃO et al., 2009). Os repositórios digitais são meios eficientes de armazenar documentos em diferentes suportes, possibilitando que uma grande quantidade de informação científica esteja disponível para acesso em rede de computadores.

O ambiente digital trouxe novos desafios para lidar com o grande volume de informações armazenadas. Especificamente no contexto dos repositórios digitais, o processo de recuperação de informação ocorre por meio de análise da sintaxe dos termos de busca inseridos pelos usuários. As principais plataformas ou softwares que operacionalizam a construção de repositórios digitais se preocupam particularmente com a descrição e com a preservação do conteúdo, utilizando modelos tradicionais de recuperação de informação. Tais características, por vezes, tem como consequência tornar o processo de recuperação de informação pouco eficiente, e por mais conteúdo que possa haver em um repositório, nem sempre as necessidades do usuário são supridas por meio de suas técnicas de busca.

Dentro desse contexto, o processo de recuperação de informação enquadra-se como uma das etapas do corpo de conhecimento da Ciência da Informação, que engloba a origem, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação (BORKO, 1968).

Os desafios propostos aos repositórios digitais não se distinguem em sua natureza das problemáticas inerentes a Web como um todo. Desde a concepção da Web, pesquisas vêm sendo realizadas buscando aprimorar a forma como a informação é tratada, desde o armazenamento até a recuperação das informações dentro desse ambiente. No entanto, a velocidade com que a difusão dessa plataforma se deu, fez com que a quantidade de dados disponibilizados crescesse de forma exponencial, tornando a organização e a recuperação processos de extrema complexidade.

Uma iniciativa que visava resolver esse problema foi proposta em 2001, por Berners-Lee, Hendler e Lassila, nomeada como Web Semântica. Essa proposta visava ser uma extensão da chamada Web atual, em que as informações seriam compreendidas tanto

por agentes computacionais, quanto pelos humanos. Em suma, a ideia dessa proposta é dar significado compreensível aos computadores, dos conteúdos disponíveis na Web. (BERNERS-LEE; HENDLER; LASSILA, 2001).

A evolução dessa proposta permitiu que estudos em diversas áreas se aproveitassem do bojo teórico e pragmático que a Web Semântica possui. Nesse sentido, os estudos de recuperação da informação estão utilizando as técnicas da Web Semântica, para tornar o entendimento das necessidades informacionais dos usuários mais claro, devido ao uso de tecnologias como ontologias, capazes de expressar formalmente aos computadores as características de um determinado domínio.

O eminente cruzamento entre os estudos de recuperação da informação, ontologias, Web Semântica e repositórios digitais será abordado nesta pesquisa, na busca de aplicar, teórica e praticamente, os conceitos que esses campos de estudos possuem, traçando algumas intersecções existentes nesse cenário.

1.1 Problema

Os repositórios digitais são meios fundamentais para a difusão do conhecimento científico; no entanto, não há, na Ciência da Informação, um número expressivo de pesquisas que se aprofundam na recuperação da informação semântica nesses ambientes. Preponderantemente, a recuperação da informação dentro dos repositórios se dá de forma sintática, analisando apenas se os termos pesquisados pelo usuário estão ou não contidos nos documentos. Como resultado desse processo, o sistema retorna uma quantidade elevada de documentos que pode não atender às necessidades dos usuários. Verifica-se que a recuperação semântica (ou recuperação por significado) não é utilizada nas mais conhecidas plataformas de repositórios digitais.

Nesse contexto, questiona-se inicialmente se o processo de busca e recuperação de informação em repositórios digitais, realizado sintaticamente, atende de maneira eficiente as necessidades informacionais do usuário. Aprofundando-se nesse cenário, identifica-se que as ontologias, juntamente com diversas outras tecnologias da Web Semântica, podem auxiliar na contextualização das informações dentro de um determinado domínio.

Embasado nessa protuberância das tecnologias e dos conceitos da Web Semântica no que tange à contextualização das informações, bem como à necessidade de haver um aprofundamento nas pesquisas que tratem da recuperação da informação em repositórios digitais, questiona-se: Como o uso de conceitos e de tecnologias da Web Semântica pode

contribuir para tornar o processo de recuperação da informação em repositórios digitais mais adequado para atender as necessidades informacionais dos usuários?

1.2 Objetivos

A pesquisa relatada nesta dissertação tem como objetivo geral propor um modelo de recuperação da informação em repositórios digitais que utilize os conceitos e as tecnologias da Web Semântica para aprimorar o entendimento da ferramenta do contexto em que as necessidades informacionais dos usuários estão inseridas. A partir desse objetivo, visa-se estender o processo de recuperação da informação com a utilização de ontologias de domínio.

Apresenta como objetivos específicos:

- Definir as relações existentes entre as propriedades do OWL e da recuperação da informação;
- Relacionar as propriedades do OWL com as consultas SPARQL;
- Definir as ações das propriedades das ontologias dentro do processo de recuperação da informação;
- Propor uma forma de interoperabilidade interativa entre repositórios digitais, em que o usuário define as fontes informacionais utilizadas;
- Implementar um protótipo que valide o modelo proposto.

1.3 Procedimentos Metodológicos

O presente estudo caracteriza-se como uma pesquisa exploratória, uma vez que foi realizada a revisão de literatura que auxiliou na construção dos conceitos da temática da pesquisa. Também se caracteriza como pesquisa descritiva, ao extrair da literatura subsídios teóricos para identificação do objeto de pesquisa, e como pesquisa aplicada, visto que foi implementado o modelo de recuperação semântica de informação aplicando-se a prova do conceito de repositórios digitais.

A execução da pesquisa ocorreu em quatro etapas:

1. Realizou-se um levantamento bibliográfico para embasar teoricamente os temas de recuperação de informação, conceitos e tecnologias da Web Semântica, ontologia, SPARQL, repositórios digitais, entre outros, bem como pesquisa de trabalhos correlatos de sistema de recuperação de informação com base em ontologias;

2. Criou-se um modelo que relaciona a recuperação da informação com as tecnologias e os conceitos de Web Semântica; a partir deste modelo;
3. Criou-se um quadro relacionando as propriedades do OWL com a recuperação de informação e, por fim, implementou-se um protótipo que liga as ontologias e os repositórios digitais;
4. Desenvolveu-se um motor de geração dinâmica de consultas SPARQL, bem como a realização de testes para verificar se o protótipo atendia aos objetivos iniciais.

As tecnologias utilizadas para o desenvolvimento desta pesquisa se apoiam na realização de buscas em sistemas do repositório, sendo que as ontologias funcionam como chave para que este processo tenha uma semântica agregada. Com isso, uma ontologia proporciona uma intersecção entre itens bibliográficos contidos nos repositórios institucionais digitais e as necessidades do usuário. Para tal, a proposta do trabalho centra-se na identificação das relações existentes entre os termos de busca inseridos pelo usuário e as ontologias, que representam o domínio da busca, onde o usuário poderá escolher as relações que satisfaçam suas necessidades informacionais. Partindo disto, o sistema conseguirá identificar as relações escolhidas e montar uma nova expressão de busca, diferenciando cada propriedade da ontologia, apresentando os elementos relevantes para especificar as necessidades informacionais do usuário.

1.4 Justificativa

A necessidade de aprofundar os estudos de recuperação da informação em repositórios digitais conduz a traçar intersecções com outros campos de estudos, como a Web Semântica. A necessidade de se contextualizarem as informações é eminente, havendo a busca por meios que possam trazer tal contextualização de forma eficiente no âmbito computacional.

As ontologias e demais tecnologias da Web Semântica permitem que mecanismos computacionais sejam capazes de compreender o sentido que as informações possuem, sendo importantes ferramentais na busca de aproximar a linguagem computacional da linguagem humana.

Dentro desse contexto, a aplicação dos conceitos e das tecnologias da Web Semântica em sistemas de recuperação da informação de repositórios digitais pode aprimorar significativamente o modo como um usuário, ao utilizar tais ambientes, recupera informações que atendam às suas necessidades informacionais. A proposta de um modelo que construa expressões de busca considerando o contexto do usuário, bem

como dos termos que compõem uma ontologia de domínio, permite que as características semânticas de um determinado contexto sejam consideradas.

O presente trabalho é embasado nessas questões, visando aprimorar a forma como um sistema compreende as necessidades informacionais dos usuários, utilizando-se de tecnologias e de conceitos oriundos da Web Semântica, para construir um modelo e uma implementação que valide as teorias propostas, buscando responder a pergunta de pesquisa explicitada anteriormente.

1.5 Trabalhos Correlatos

Há uma série de estudos que visam utilizar ontologias no contexto da recuperação da informação nos mais diversos domínios. Nesta subseção serão apresentados alguns desses trabalhos, que foram escolhidos pela proximidade dos objetivos, pela quantidade de citações ou pela relevância dentro da área da Ciência da Informação.

Um primeiro trabalho de destaque, com um alto número de citações é o de Jones, Alani e Thudope (2001), em que os autores promovem a recuperação da informação de dados geográficos. Nesse trabalho, os autores, além de criar uma forma de se recuperar as informações, constroem uma ontologia para atender o seu domínio. O trabalho desenvolvido lida com cálculos matemáticos que relacionam algumas informações com os dados geográficos, buscando melhorar a relevância das informações que são obtidas, por meio do cruzamento das coordenadas, bem como da proximidade em relação a temas de interesses.

Esse primeiro trabalho, com forte ligação com a Ciência da Computação, apresenta objetivos semelhantes a esta pesquisa, uma vez que há a necessidade de melhorar a relevância na recuperação da informação. No entanto, a ontologia construída não apresenta uma semântica formal próxima a que o OWL fornece. Vale destacar que no ano de publicação desse trabalho não havia a linguagem OWL, o que faz com que haja uma diferença significativa na capacidade de descrição obtida entre a ontologia utilizada nesse trabalho e no OWL. Além disso, o trabalho apresentado pesquisa sobre os resultados obtidos, diferenciando-se desta pesquisa, uma vez que trabalhamos sobre a expressão de busca, que refletirá a semântica extraída da ontologia.

Já Janaite Neto e Ferneda (2016) apresentam uma proposta em que as ontologias são utilizadas como ferramental de apoio à indexação de documentos, para a inserção de

termos relacionados, além do uso da ontologia para a expansão da busca do usuário, de acordo com as terminologias da ontologia.

Uma parte do trabalho apresentado apresenta relação com a proposta desta pesquisa, no que tange à inserção de novos termos de expressão na busca do usuário. Porém, difere ao propor a indexação dos documentos, objetivo que não abrangemos neste trabalho. Desta feita, limitamo-nos aqui a discutir as relações referentes a questões relativas à expansão da busca do usuário.

Os autores relatam que após o usuário escrever uma expressão de busca, os termos definidos serão buscados na ontologia, sendo que após sua localização, caso o termo seja encontrado, serão considerados os termos genéricos e específicos oriundos da ontologia, como uma estrutura taxonômica. O principal diferencial do presente trabalho diz respeito à utilização das ontologias, não somente como estruturas de taxonomia, mas utilizando, essencialmente, as relações semânticas que apresentam uma contextualização formal dos termos localizados. Assim, as diferenças dos trabalhos são que neste há uma expansão na utilização das propriedades de ontologias, mais especificamente do OWL, enquanto que no trabalho de Janaíte Neto e Ferneda (2016) existe uma integração com a indexação dos termos e o uso do Modelo Vetorial, não sendo esses objetivos tratados no presente trabalho.

Em um artigo dentro do contexto da Ciência da Computação, Jain e Singh (2013) fazem uma revisão bibliográfica dos artigos que tratam de recuperação da informação baseada em ontologias. Nesse contexto, as pesquisas que tratam de expansão de expressões de buscas são aquelas que mais se aproximam dos objetivos propostos no presente trabalho. Os autores consideram que os objetivos desses trabalhos estão vinculados à relevância dos termos escolhidos pelos usuários, que poderão ou não ser relevantes de acordo com cada busca; a partir disso, são aplicadas algumas ontologias que visam criar consultas separadas, baseadas nas regras das ontologias, e dessa forma, criar uma nova expressão de busca.

Há uma profunda relação entre esses processos de expansão de busca com os processos realizados no presente trabalho. Essa relação se dá principalmente pelas modificações na expressão de busca, de acordo com as relações entre os termos. No entanto, a principal diferença entre estes trabalhos com a presente pesquisa está na utilização e definição das propriedades do OWL, como principal ferramental na definição da expressão de busca, em que o processo de modificação (ou expansão) de busca se dá baseado em ações relacionadas a cada propriedade. Um outro ponto que diferencia o

presente trabalho é o fato de haver uma série de conceitos e de tecnologias da Web Semântica trabalhando para a modificação da expressão de busca, especialmente o Motor de SPARQL, que utiliza o SPARQL e as propriedades do OWL para explorar e localizar as relações existentes na ontologia.

Dessa forma, a semântica do processo ocorre tanto ao explorar a ontologia, quanto na montagem da nova expressão de busca. Os pontos relatados podem ser identificados no trabalho de Shah, Finin e Joshi (2002).

Neste trabalho, se tomar como referência a posição desses autores, pode-se apontar uma expansão realizada, utilizando a linguagem DAML+OIL, havendo o uso de inferências juntamente com as ontologias. Novamente neste trabalho, o uso das ontologias analisa as relações, porém sem um aprofundamento nas propriedades que as ontologias possuem, de modo a não haver uma ação específica para cada tipo de propriedade existente.

Por meio dos trabalhos apontados, identificam-se as contribuições que a presente pesquisa apresenta, além de serem demonstradas as principais diferenças entre esses trabalhos, de importância para as áreas da Ciência da Informação e da Ciência da Computação, e o trabalho desenvolvido e exposto neste texto.

1.6 Estrutura do Trabalho

O trabalho está dividido em oito seções. Na primeira seção é apresentado a introdução, contendo o problema, os objetivos, os procedimentos metodológicos, a justificativa, os trabalhos correlatos e a estrutura do trabalho.

Na sequência, as seções dois, três e quatro contemplam as discussões teóricas em que o trabalho se sustenta, apresentando respectivamente repositórios digitais, web semântica e recuperação da informação.

Na seção cinco é apresentado o modelo de Recuperação da Informação para Repositórios baseado em ontologias, dividindo a seção nas camadas do modelo: Apresentação, Controle, Semântica e Dados. Já na seção seis demonstra-se como se deu a implementação do protótipo, dividindo a seção igualmente nas camadas do modelo.

Por fim, a seção sete apresenta os trabalhos futuros que podem ser realizados e a seção oito contempla as considerações finais. Ao final são apresentadas as referências.

2 REPOSITÓRIOS DIGITAIS

A explosão da geração informacional, ocorrida em maior escala no início do século XXI, proporcionou uma revolução no modo como os documentos são gerados e disponibilizados. Entre os principais agentes desse fenômeno, destaca-se a Web e a Internet, que propiciou que os documentos digitais se difundissem nos diferentes contextos da sociedade, passando desde a geração de notas fiscais até a submissão de um artigo em um periódico científico, processos que ocorrem majoritariamente em ambientes digitais.

A ocorrência desses processos em ambientes digitais proporcionou que importantes passos fossem dados dentro do contexto científico. Uma eminente consequência disto, foi o aumento que a produção científica teve nas últimas décadas, ocorrendo uma expansão significativa na geração de materiais científicos (SANTAREM SEGUNDO, 2010).

Esse crescimento na produção ocorre, entre outros fatores, por uma facilidade que as Tecnologias de Informação e Comunicação proporcionam, de acesso aos trabalhos científicos, pelo uso de comunicadores instantâneos, em que pessoas espalhadas por todo globo podem se comunicar; ferramentas de produções de textos em conjunto, permitindo que pesquisadores produzam conhecimento científico cooperativamente; e plataformas para a disponibilização de documentos, que podem ser acessados abertamente por qualquer usuário conectado. Dessa forma, a colaboração científica passou a ser uma realidade cada vez mais presente no contexto dos pesquisadores, em que, com certa facilidade, acadêmicos de distintas instituições, em países distantes, podem produzir ciência em colaboração.

Um fator fundamental nesses processos, é o uso de formatos digitais para que usuários possam acessar os documentos, em qualquer local em que estejam conectados à rede, tornando o compartilhamento das informações uma tarefa cada vez mais necessária e comum (CATARINO; BAPTISTA, 2007).

No âmbito acadêmico, paralelo ao crescimento das publicações científicas, houve a necessidade de que as informações estivessem organizadas, para que outros pesquisadores pudessem acessar os conhecimentos produzidos, permitindo, assim, uma maior colaboração entre os pares, com acesso a outros trabalhos, e evitando também que pesquisadores desenvolvessem pesquisas com objetivos semelhantes, e que elas

pudessem, isso sim, acrescentar avanços umas às outras, permitindo uma evolução na ciência como um todo.

Entretanto, pelo número crescente de produções e a organização ineficiente da Web, ocorreu uma carência de tecnologias que aprimorasse o uso desta plataforma no meio acadêmico. Na busca de uma solução para este problema, os repositórios digitais foram criados para armazenar as produções científicas. Repositório digital é definido, pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), como base de dados que reúne, de uma maneira organizada, a produção científica de uma instituição ou de uma área temática (IBICT, 2015).

Uma definição mais completa, que abrange outros aspectos, traz que os repositórios digitais podem ser compreendidos como “[...] uma forma de armazenamento de objetos digitais que tem a capacidade de manter e gerenciar material por longos períodos de tempo e prover o acesso apropriado”. (VIANA, ARELLANO, SHINTAKU, 2013).

Tais definições foram sendo aprimoradas, seguindo uma evolução dos repositórios digitais, que deixaram de ser simples sistemas de armazenamento de produções científicas e passaram a ser um ambiente que promove diversas interações entre usuário e sistema, perpassando por questões como colaboração e facilidade na comunicação entre pesquisadores, deixando de ser um sistema de informação estático, em que o usuário simplesmente coleta o documento que ele necessita (SAYÃO et al., 2009).

Essas definições trazem diversos aspectos envolvidos em repositórios digitais, tais como armazenamento de documentos, a preservação digital, a interação entre usuários e sistemas, a utilização de metadados, entre outros fatores, que são fundamentais para o entendimento dos repositórios digitais. Em síntese, os repositórios digitais podem ser entendidos como sistemas de informação que promovem o armazenamento e a disponibilização de objetos digitais, favorecendo a interação entre os usuários e o sistema, na busca de permitir um aprimoramento da recuperação de informação dentro do ambiente.

Um exemplo da contribuição dos repositórios digitais pode ser observado na disseminação de trabalhos de monografia, de dissertações e de teses, permitindo que cada vez mais estudantes de diversas instituições tenham acesso a essa fonte de informação. Deve-se destacar a importância desse fato, tendo em vista que essas fontes de informações são fundamentais para o desenvolvimento da maturidade e da experiência científica iniciante dos acadêmicos (PETINARI, 2008).

Para uma melhor compreensão dos usos que os repositórios digitais possuem, podemos classificá-los em três grupos: repositórios institucionais, focados em reunir a produção intelectual de uma instituição; repositórios temáticos ou disciplinares, que buscam reunir a produção de áreas do conhecimento específicas, voltadas para determinadas comunidades científicas; e repositórios de teses e dissertações, que reúnem produções desses tipos (LEITE et al., 2009).

Um grupo de repositório que está ganhando grande importância, por conta da alta adesão das instituições, é o dos repositórios institucionais. Tal tipo de repositório tem como premissa reunir a produção intelectual e científica de uma determinada instituição, de forma que, além de um trabalho estar disponível no ambiente digital do periódico onde foi publicado, esse mesmo trabalho estará acessível também em um repositório da instituição, em que os pesquisadores estejam vinculados, sem deixar de cumprir com as questões legais envolvidas.

Shintaku (2015, p. 16) afirma que um “[...] ponto importante que caracteriza o repositório institucional é a vinculação de pelo menos um dos autores à instituição responsável pelo repositório.” Dessa forma, o autor trata diretamente da função do repositório institucional, em que existe um vínculo direto entre as produções criadas pelos autores pertencentes à instituição daquele repositório.

Os repositórios institucionais popularizaram-se entre organizações e instituições, na busca de proporcionar o livre acesso às produções científicas, promovida principalmente com uma iniciativa chamada *Open Access Initiative* (OAI), que busca suscitar o acesso livre e irrestrito à literatura científica e acadêmica (SANTAREM SEGUNDO, 2010).

Assim, o uso de repositórios abertos permite que muitos conteúdos possam ser acessados sem custos e sem barreiras, oportunizando que os resultados de pesquisas possam ser mais amplamente difundidos, estimulando, assim, a pesquisa científica como um todo (LEITE; ARELLANO; MORENO, 2006).

Para possuir um melhor panorama do uso de repositórios digitais, um serviço, chamado de *Registry of Open Access Repositories*¹ (ROAR), realiza a verificação das informações de diversos repositórios espalhados pelo mundo. Em levantamento realizado em novembro de 2015, pode-se observar que dos 4098 repositórios cadastrados nesse

¹ Registry of Open Access Repositories. Disponível em: <<http://roar.eprints.org/>> Acesso em: 30 nov. 2015.

sistema, 2772 eram repositórios institucionais, correspondendo a quase 68% de todos os repositórios. Tal dado mostra claramente o interesse que as instituições vêm demonstrando em criar meios de divulgarem e armazenarem suas produções científicas, além de evidenciar que os repositórios institucionais representam a principal atribuição que as implementações de repositórios digitais possuem no momento.

Cabe uma observação aos dados apresentados, em que os números fornecidos podem não representar fielmente o número total de repositórios digitais espalhados pelo mundo, pois o cadastro dos repositórios nesse sistema é voluntário, como relata Shintaku (2015). O autor relata ainda que em fevereiro de 2014 existia um quantitativo de 3478 repositórios institucionais cadastrados no ROAR, mostrando que em 22 meses ocorreu um aumento de 620 repositórios digitais cadastrados, o que representa um aumento de aproximadamente 18% no número de repositórios.

A partir desses resultados, é possível visualizar que o uso de repositórios digitais vem aumentando substancialmente com o passar dos anos, sendo um meio eficiente de divulgar produções científicas, além de ser uma fonte de informação muito valiosa para as instituições e para grupos científicos, pois promove o livre acesso à informação.

Um histórico mais aprofundado da evolução dos repositórios digitais foi realizado por Santarem Segundo (2010), abordando a forma como os repositórios foram concebidos, bem como o momento histórico vivido na época. O presente trabalho não se aprofunda em tais questões, pois não é objetivo desta pesquisa, que utiliza os repositórios como um meio de demonstrar técnicas e teorias de recuperação da informação e de Web Semântica.

Um dos principais componentes dos repositórios digitais são os metadados, que são utilizados para descrever os documentos que estão armazenados dentro dos repositórios digitais. Dessa forma, os metadados representam um papel central nesta pesquisa, por serem o objeto de busca em que será realizada a recuperação da informação. A seguir, discorre-se acerca dos metadados.

2.1 Metadados

O aumento da quantidade de documentos disponíveis nos meios tradicionais e nos meios digitais, está desafiando a Ciência da Informação, principalmente as disciplinas de catalogação e de recuperação da informação. Nesse contexto, a consolidação dos processos de catalogação foi marcada pelo desenvolvimento de regras e diversos códigos,

ao mesmo tempo em que se utilizou das tecnologias disponíveis na época para agilizar esses procedimentos.

Um formato que apresenta grande importância nesse cenário é o formato automatizado *Machine Readable Cataloging* (MARC). Este formato passou por diversas atualizações e revisões até ser lançada a versão que é utilizada atualmente, denominada de MARC21, tendo uma grande importância para o contexto bibliográfico, bem como nos estudos de metadados.

A partir dos anos 90, o termo metadados começou a ser amplamente utilizado e começaram a ser desenvolvidos padrões de metadados para a descrição de recursos. A aplicação do termo metadado, a princípio, ocorreu dentro da área da Ciência da Computação, para especificar objetos de bancos de dados. Posteriormente, em meados da década de 1990, tal termo passou a ser empregado para a descrição de recursos na Web, com o desenvolvimento do padrão *Dublin Core*. O *Dublin Core* surgiu no EUA, em março de 1995, como um conjunto mínimo para descrever informações disponíveis na Internet. (ALVES; SANTOS, 2013).

O termo metadados apresenta o significado literal de “dados sobre dados”, ou seja, dados que dizem respeito a outros. Contudo, tal definição se mostra bastante genérica, pois é “[...] uma definição que pouco explica seu significado e não resolve o problema da pluralidade discursiva do conceito de metadados.” (ALVES; SANTOS, 2013, p. 38).

Desta forma, Alves (2010, p. 47) define metadados como:

[...] 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.

Uma outra definição que aborda diversos aspectos dos metadados é dada por Grácio (2002, p. 21):

Comumente chamado de dados sobre dados, o termo metadados pode ser mais bem descrito como um conjunto de dados chamados de elementos, cujo número é variável de acordo com o padrão, e que descreve o conteúdo de um recurso, possibilitando a um usuário ou a um mecanismo de busca acessar e recuperar esse recurso. Esses elementos descrevem informações como nome, descrição, localização, formato, entre outras, que possibilitam um número maior de campos para pesquisas.

A partir das definições dada por esses autores, identifica-se uma relação intrínseca entre os metadados e a descrição de recursos, em que os metadados se mostram

fundamentais na representação de objetos do mundo real. O conceito de metadados está relacionado com o conceito de padrões de metadados, que pode ser compreendido a partir da explanação feita por Alves e Santos (2013, p. 43): “[...] os metadados (*metadata*) devem ser codificados em estruturas padronizadas de descrição, denominadas como padrões de metadados (*metadata statement*).”

Os padrões de metadados foram criados com o intuito de formalizar e padronizar os metadados. Rosseto (2003) relata que os padrões de metadados “[...] estabelecem regras para a definição de atributos (metadados) de recursos informacionais [...]”. Alves (2010, p. 47-48) complementa a definição, dizendo que “O objetivo do padrão de metadados é descrever uma entidade gerando uma representação unívoca e padronizada que possa ser utilizada para recuperação da mesma.”

Assim, compreende-se que os metadados são os atributos que descrevem e caracterizam uma entidade ou objeto, enquanto o padrão de metadados é o conjunto de metadados. Este último pode comportar regras para a descrição de alguma informação em um metadado, o que depende do domínio em questão. Outra característica, que depende do padrão de metadados, é a existência de instruções do próprio esquema ou a indicação de regras e códigos externos para a representação mais específica dos objetos. Verifica-se que os padrões de metadados são de grande importância para que os recursos possam ser descritos adequadamente para o domínio ao qual eles pertencem.

No contexto do domínio bibliográfico, o principal padrão de metadados é o formato automatizado MARC21. Este formato foi desenvolvido por instituições e profissionais com vasta experiência, para realizar a descrição dos itens bibliográficos, atendendo aos requisitos e às necessidades do domínio bibliográfico.

Os elementos do MARC 21 estão dispostos em campos (*tags*) e subcampos representados por números e letras. Cada campo (metadado) descreve uma informação ou um conjunto de informações bibliográficas sobre diferentes tipos de materiais. A sintaxe para armazenamento e intercâmbio dos dados MARC 21 foi baseada na norma ISO 2709, que apresenta uma estrutura linear de armazenamento dos dados. Destaca-se que essa estrutura do ISO 2709 surgiu para o intercâmbio de dados por meio de fitas magnéticas, na década de 1960, o que tornou o padrão MARC21 bastante dependente da codificação (ASSUMPCÃO; SANTOS, 2015; FUSCO, 2010).

Com a Web, alguns questionamentos têm sido recorrentes sobre o uso do MARC21 nos dias atuais. Contudo, deve-se ressaltar a necessidade de ter cautela ao enunciar o MARC como obsoleto, pois este tem atendido as necessidades de descrição e

intercâmbio entre bibliotecas, por meio de metadados específicos para a catalogação e recuperação de objetos bibliográficos. Destaca-se aqui a necessidade de adaptação da sintaxe dos dados bibliográficos do padrão MARC 21 em uma estrutura compatível com a Web.

Elencadas essas questões, e a partir delas, é importante ressaltar as diversas discussões que estão ocorrendo, principalmente no que se refere à sintaxe do MARC21, a ISO 2709, pois ela não favorece a interoperabilidade na Web, diferentemente dos padrões de metadados baseados na sintaxe - *eXtensible Markup Language* (XML).

O MARC 21 possui uma versão em XML, denominada MARCXML; entretanto, mesmo sendo expresso em uma sintaxe adequada para a Web, os metadados (campos e subcampos do MARC21) ainda são representados em norma de números e letras, dificultando sua compreensão. Além disso, o MARCXML apresenta a mesma especificidade do que o MARC 21, o que muitas vezes o torna complexo para ser utilizado na Web.

No âmbito da Web, o padrão de metadados *Dublin Core* tem ganhado um destaque significativo, havendo uma ampla gama de produções científicas tratando desse padrão. O padrão *Dublin Core* é formado por quinze elementos, tendo como objetivo facilitar a identificação e descoberta de recursos na Web; portanto, caracteriza-se como um padrão para propósitos gerais, ou seja, pode ser utilizado em diversos domínios. O quadro 1 apresenta os quinze elementos do *Dublin Core*, com uma breve descrição de cada elemento.

Quadro 1 - Elementos do *Dublin Core*

NOME	DESCRIÇÃO
<i>Contributor</i>	Uma entidade responsável por realizar contribuições para o recurso.
<i>Coverage</i>	O tópico espacial ou temporal do recurso, a aplicabilidade espacial do recurso ou da jurisdição sob a qual o recurso é relevante.
<i>Creator</i>	Uma entidade principal responsável por fazer um recurso.
<i>Date</i>	Um ponto ou período de tempo associados com um evento no ciclo de vida do recurso.
<i>Description</i>	Uma conta do recurso.
<i>Format</i>	O formato de arquivo, meio físico, ou dimensões do recurso.

<i>Identifier</i>	Uma referência não ambígua ao recurso dentro de um determinado contexto.
<i>Language</i>	O idioma do recurso.
<i>Publisher</i>	Uma entidade responsável por tornar o recurso disponível.
<i>Relation</i>	Um recurso relacionado
<i>Rights</i>	Informações sobre os direitos de e sobre o recurso.
<i>Source</i>	Um recurso relacionado a partir do qual o recurso descrito é derivado.
<i>Subject</i>	O assunto do recurso
<i>Title</i>	Um nome dado a um recurso
<i>Type</i>	A natureza ou gênero de um recurso

Fonte: DCMI, 2012, tradução nossa.

No documento (DCMI, 2012), que descreve os quinze elementos apontados pelo quadro 1, é informado que tais elementos fazem parte de um conjunto maior de termos mantidos pelo *Dublin Core Metadata Initiative* (DCMI), chamado de *DCMI-Terms*. Além disso, o padrão *Dublin Core* passa por constantes melhorias, para que possa se adaptar aos novos contextos da Web, como a própria Web Semântica.

Outro ponto de destaque quando se trata dos quinze elementos do *Dublin Core*, é que eles são formalmente aprovados nos padrões de referência *International Organization for Standardization* (ISO 15836:2009), *American National Standards Institute / The National Information Standards Organization* (ANSI/NISO Z39.85-2012) e *Internet Engineering Task Force Request for Comments* (IETF RFC 5013) (DCMI, 2012). A aprovação formal por parte desses órgãos concede ao *Dublin Core* uma maior confiança em seu uso, ao demonstrar que existe uma padronização, já que instituições renomadas garantem isso.

Outra característica que insere uma maior capacidade de descrição no padrão *Dublin Core*, é a possibilidade da utilização de qualificadores. Tais qualificadores permitem que seja inserido uma maior especificidade nos quinze elementos, sendo adequados em alguns contextos, em que é necessária uma maior granularidade na descrição dos recursos informacionais (DCMI, 2000).

Além do *Dublin Core*, existem diversos outros padrões, tanto para propósitos específicos quanto para propósitos gerais, e conforme os requisitos do domínio devem ser

adotados padrões que os satisfaçam. Todavia, no contexto de repositórios digitais, existe um uso majoritário do *Dublin Core* como estrutura de representação dos registros. Tal fato ocorre pela aderência que existe entre as necessidades apresentadas pelos repositórios digitais e os recursos oferecidos pelo *Dublin Core*, pois repositórios digitais, em suma, necessitam de uma representação adequada, para que tornem possível a recuperação de informação, porém não necessita de uma descrição tão detalhada quanto a fornecida em padrões como o MARC21.

Outro ponto que relaciona diretamente o uso do *Dublin Core* é a sua utilização como padrão de metadados inicial do software de repositórios digitais DSpace, como relata Viana, Arellao e Shintaku (2013). Os autores complementam dizendo ainda que o DSpace permite o uso dos quinze elementos do *Dublin Core*, mas também possibilita o uso em associação com os qualificadores do padrão, que irão inserir maior especificidade aos quinze elementos principais. Dessa forma, como o *DSpace* é um dos softwares de repositórios digitais mais utilizados no momento, e o *Dublin Core* consegue atender as necessidades desse domínio, tal padrão apresenta grande importância no contexto de repositórios digitais.

A utilização de um padrão de metadados permite que possa ocorrer interoperabilidade entre distintos repositórios digitais. Nesse âmbito, a interoperabilidade entre objetos informacionais digitais em repositórios digitais está fortemente vinculada ao uso do protocolo *Open Archives Initiative - Protocol for Metadata Harvesting* (OAI-PMH). A seguir, será descrito com maiores detalhes o protocolo OAI-PMH.

2.2 Protocolo OAI-PMH

A iniciativa *Open Archives* (Arquivos Abertos) surgiu nos Estados Unidos da América em 1999, com o objetivo de desenvolver formas de interoperabilidade e de promover a disseminação mais eficiente das informações (LAGOZE; VAN DE SOMPEL, 2001). Interoperabilidade, dentro da área de tecnologias da informação, relaciona-se com a comunicação entre programas de computadores e, mais especificamente, pode ser compreendida como um processo para garantir que os sistemas, os procedimentos e a cultura de uma organização possam ser gerenciados, permitindo a maximização das oportunidades para troca de dados (CASTRO, SANTOS, 2014).

Embasado nessas afirmações, verifica-se a importância que a iniciativa do protocolo OAI-PMH possui ao permitir o compartilhamento de metadados. A versão inicial do OAI-PMH foi lançada em 2001, existindo uma versão 2.0, que é a última versão

lançada, datada de julho de 2002 (GARCIA; SUNYE, 2003). Tal protocolo utiliza o *Dublin Core* como padrão de metadados para possibilitar o intercâmbio das informações, permitindo, assim, que a interoperabilidade possa ocorrer dentro dos ambientes atendidos.

Por meio do uso do protocolo OAI-PMH, um sistema pode obter os metadados dos objetos que estão contidos em um repositório. Para tal, basta realizar uma requisição para o endereço onde o repositório se encontra, além de inserir as configurações daquela coleta. Como resultado, é retornado uma lista que contém todos os metadados daquele repositório.

Um benefício propiciado pelo protocolo OAI-PMH é a facilidade com que ocorre a distribuição das informações científicas, que, por meio da interoperabilidade proposta, permite que exista maior visibilidade e facilidade no acesso às produções científicas contidas em periódicos, anais de eventos, capítulos de livros, teses e dissertações. Dessa forma, um editor de uma revista, por exemplo, consegue coletar os registros de objetos produzidos em sua revista e que estão espalhados em diversos locais, pois os dados estão abertos, permitindo o compartilhamento entre as organizações. Outro exemplo muito utilizado é o de instituições de ensino, que podem obter, por meio do OAI-PMH, o acesso às produções intelectuais que estão espalhadas por distintos repositórios digitais (VIANA; ARELLANO; SHINTAKU, 2013).

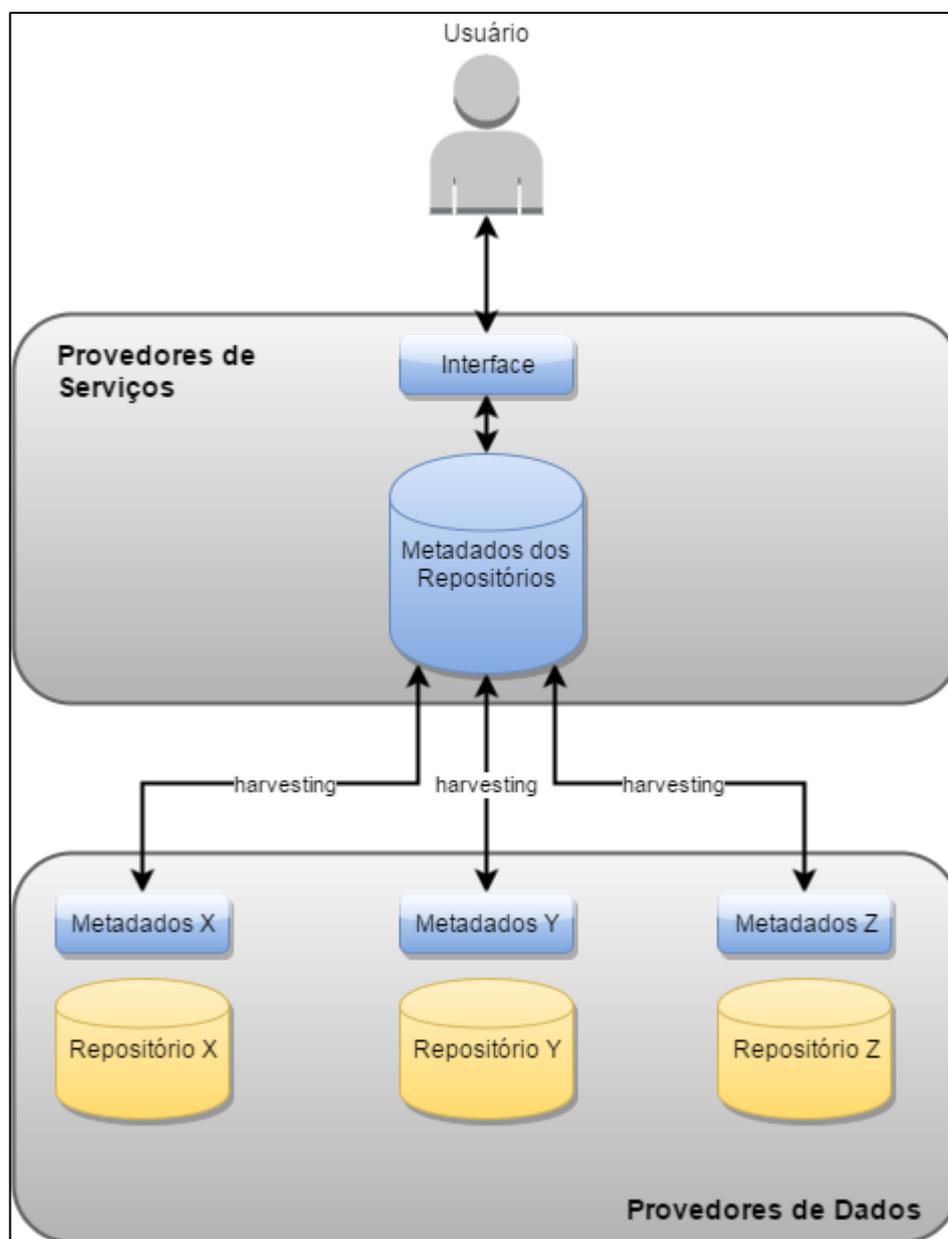
Dentro do contexto de múltiplos repositórios, existem basicamente dois atores principais que utilizam o OAI-PMH: os provedores de dados e os provedores de serviços. Os provedores de dados são as bases de dados onde os objetos digitais estão disponíveis para serem acessados. Essas bases de dados, caso utilizem o protocolo OAI-PMH, permitem que seus metadados possam ser coletados por meio de um programa de *harvesting*. O segundo ator do processo, os provedores de serviços, realizam o processo de coleta dos metadados nos diversos provedores de dados. Essa coleta pode ser compreendida como uma interface entre os usuários e os metadados das bases de dados, pois é por meio do provedor de serviço que o usuário conseguirá localizar os documentos, em plataformas que utilizam o protocolo OAI-PMH (MARCONDES; SAYÃO, 2001).

Como relatado, a coleta de dados realizada com o protocolo OAI-PMH, chama-se *harvesting*. Compreende-se *harvesting* como:

[...] um processo unilateral, onde, os provedores de serviços, a partir da lista de repositórios (provedores de dados) registrados na OAI, realizam periodicamente uma busca a estes provedores de dados, "colhendo" os metadados para exibição sob a forma de consultas efetuadas pelos usuários (GARCIA; SUNYE, 2003, p. 4).

Uma arquitetura padrão de interoperabilidade com o uso do protocolo OAI-PMH pode ser visualizada na Figura 1, em que é possível verificar como ocorre a relação entre os provedores de dados e os provedores de serviço.

Figura 1 - Arquitetura Padrão do Protocolo OAI-MH



Fonte: Elaborado pelo autor

Por meio da Figura 1, visualiza-se como ocorre a interação entre os provedores de dados e de serviços. É possível visualizar que o processo de *harvesting* realiza a coleta dos metadados nos provedores de dados, passando para uma base de dados que contém os metadados coletados. Outro ponto a ressaltar é a interação do usuário com o sistema,

que ocorre em uma interface que possibilita o acesso do usuário à base de dados dos metadados.

Destaca-se ainda que os provedores de serviço irão coletar apenas os descritores dos arquivos; por tal motivo, os repositórios fazem parte dessa arquitetura, pois são eles que armazenam os objetos digitais. Vale ressaltar, também, que a figura contém três repositórios (repositório X, repositório Y e repositório Z), entretanto uma implementação do esquema de interoperabilidade em repositórios digitais, pode conter um ou inúmeros repositórios, sendo utilizado três, apenas, como mecanismo de ilustração.

3 WEB SEMÂNTICA

A World Wide Web foi criada em 1989 com o objetivo de permitir o acesso, o intercâmbio e a recuperação de informação. No texto chamado de *Information Management: A Proposal*, de 1989, Berners-Lee inicia os debates acerca de novos meios de disponibilizar a informação, refletindo sobre as estruturas mais adequadas para armazenar e permitir a navegação dos usuários (BERNERS-LEE, 1989).

Passados mais de 25 anos da criação da Web, verifica-se uma presença maciça do hipertexto em todo esse ambiente, em que novos conceitos e tecnologias, tais como mídias sociais, *wikis*, *ebooks*, repositórios e bibliotecas digitais, utilizam o hipertexto como um meio para promover interatividade entre os usuários e entre os usuários e os sistemas.

Com o hipertexto e todas as opções existentes, a Web apresentou um crescimento exponencial após o seu surgimento, sem, todavia, contar com uma organização adequada para a quantidade de dados que estava disponível para o acesso. Outra característica de destaque no modo como o crescimento da Web ocorreu, foi que as páginas e os dados presentes estavam preparados, em sua maioria, unicamente para leitura humana, não apresentando expressividade para que agentes computacionais compreendessem o sentido das informações (SOUZA; ALVARENGA, 2004).

Essa característica - de disponibilizar informações somente para leitura humana – foi contrária a proposta inicial da Web, de Berners-Lee, que discorria a respeito da necessidade de que, além de pessoas, as máquinas também fossem capazes de ler os dados e trabalhem com eles.

Souza e Alvarenga (2004, p. 133) demonstram a situação vivida pela Web nos primeiros anos após sua criação:

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.

Essa situação descrita pelos autores, por mais que demonstre um momento anterior ao vivido atualmente, mostra com clareza os problemas e as características que a Web apresentou, principalmente em seu início, mas que ainda persistem, de alguma forma.

Com a intenção de aprimorar essa questão e criar ambientes onde a interação entre humanos e máquinas ocorresse com maior facilidade, Berners-Lee, Hendler e Lassila

(2001) propuseram em 2001 a Web Semântica. Nesse texto, chamado de *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, os autores expõem como a Web deveria trabalhar para possibilitar que agentes computacionais pudessem trabalhar em cooperação com as pessoas, em que os usuários não necessitariam coletar os dados em diversos Web Sites distintos para conseguir obter as informações que atendessem às suas necessidades informacionais. Nessa proposta, os agentes computacionais iriam se comunicar e conseguir entender o significado dos dados para entregar com eficiência aquilo que interessa ao usuário. (BERNERS-LEE; HENDLER; LASSILA, 2001).

Para promover isso, os autores trazem a necessidade da aplicação e da construção de novos conceitos e tecnologias que trabalhariam em conjunto e tornariam real a proposta feita. Outro ponto destacado, é que os dados deveriam estar interligados e serem interoperáveis, promovendo, assim, um fácil acesso às informações (BERNERS-LEE; HENDLER; LASSILA, 2001).

Berners-Lee, Hendler e Lassila (2001, tradução nossa) descreveram 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. Os primeiros passos para tecer a Web Semântica a partir da estrutura da Web existente já estão a caminho. Em um futuro próximo, este desenvolvimento irá inaugurar novas funcionalidades significativas, de forma que as máquinas se tornarão muito mais capazes de processar e "entender" os dados do que simplesmente exibi-los.

Dessa forma, essa extensão da Web permitirá que os conteúdos sejam compreendidos e gerenciados, independentemente do gênero a que ele pertencer (música, vídeos, publicações de blogs, *wikis*), por meio do valor semântico que tais dados possuem.

Uma outra contextualização realizada, tendo como base a Web Semântica, é dada pelo *World Wide Web Consortium (W3C)* (consórcio que administra a Web), a respeito da Web de documentos e a Web de dados. A Web de documentos corresponde a Web tradicional, em que os usuários acessam documentos construídos principalmente para a leitura humana. A Web de dados corresponde a uma evolução da Web de documentos, onde seria possível realizar buscas como em um banco de dados, tendo como objetivo final permitir que os computadores realizem tarefas mais úteis para os usuários (W3C, 2011).

A W3C ainda complementa essa visão de Web de dados, inserindo algumas funções que a Web Semântica deve atender, como permitir que pessoas possam criar repositórios de dados na Web, construir vocabulários e escrever regras que permitam a interoperabilidade entre esses dados (W3C, 2011).

Possuindo uma visão semelhante à apresentada pela W3C, Karin Breitman (2005) insere os conceitos de Web Sintática e Web Semântica. A autora discorre sobre a Web Sintática, dizendo que a Internet atual apresenta características singulares, em que o computador fica responsável essencialmente pela organização da informação, e fica a cargo dos seres humanos realizarem a interpretação de toda a informação.

Breitman (2005) ainda diz que tal problema ocorre, pois principalmente as páginas Web não contêm, em sua maioria, informações que descrevem o conteúdo ali presente. A autora complementa refletindo sobre a maior dificuldade encontrada, como consequência a essa característica da Web, dizendo que é a ineficiência dos mecanismos de buscas, que, em suma, recuperam uma grande quantidade de páginas, que podem até apresentar documentos que contemplem as necessidades informacionais dos usuários, porém estão em meio a milhares de outras páginas que eles podem não chegar a contemplar.

Ramalho, Vidotti e Fujita (2007), complementam tais questões, abordando a relação entre as máquinas e os seres humanos, dentro do âmbito da Web:

[...] 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.

Ao relacionar o papel das máquinas com os humanos dentro da Web Semântica, os autores demonstram a necessidade de haver uma melhor comunicação entre esses atores, tendo uma relação direta com artefatos capazes de fornecer sentido aos mecanismos computacionais.

Enfocando os desafios que a Web Semântica deveria percorrer para resolver as questões elencadas, Souza e Alvarenga (2004, p. 134) definem o projeto da Web Semântica tendo um olhar focado em como essa proposta poderia se tornar real.

O projeto da Web Semântica, em sua essência, é a criação e implantação de padrões (*standards*) tecnológicos para permitir este panorama, que não somente facilite as trocas de informações entre agentes pessoais, mas principalmente estabeleça uma língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

Essa definição dada por Souza e Alvarenga (2004) demonstra que a Web Semântica está relacionada diretamente a panoramas tecnológicos, que irão promover melhorias e avanços no modo como a Web se organizará, além de estabelecer como ocorrerá o processo de comunicação entre agentes e usuários. Nesse sentido, Santarem Segundo (2010, p. 72) aprofunda tais questões, elencando o ponto fundamental da Web Semântica:

O projeto da Web Semântica tem como ponto fundamental a criação de uma nova estrutura de armazenamento de dados. O ponto principal está na separação da apresentação do conteúdo e do conteúdo da estrutura, tratando as unidades atômicas de uma informação como componentes independentes.

Uma importante característica que revela a necessidade de a Web Semântica migrar da Web atual, é a questão da descentralização desmedida, uma vez que não há nenhuma organização ou governo responsável pela Web. Tal fato apresenta-se como um desafio, na busca de organizar a informação, pois a descentralização pode promover a desorganização, necessitando existir o compromisso de haver meios para tornar possível uma organização adequada. Uma segunda solução possível, seria a existência de diversos modelos de organização em cada instituição, sendo que a descentralização estaria contribuindo na promoção de meios de organização (BREITMAN, 2005).

Posteriormente a sua ideia inicial, a Web Semântica vem se difundindo, e diversas tecnologias estão sendo implementadas. Entre as diversas tecnologias destacam-se o *Resource Description Framework* (RDF), a *Web Ontology Language* (OWL), a *eXtensible Markup Language* (XML), o *Uniform Resource Identifier* (URI), e diversas outras tecnologias e conceitos que são descritos pelo W3C.

Santarem Segundo (2014) afirma que tais tecnologias tornam possível a materialização da Web Semântica, estando relacionadas principalmente ao processo de armazenamento e de construção das informações. A seguir serão abordadas mais detalhadamente as tecnologias base da Web Semântica.

3.1 Conceitos e Tecnologias da Web Semântica

Como relatado nas definições de alguns autores sobre a Web Semântica, esta é composta por diversas tecnologias que buscam torna-la real e implementável. Ferreira (2014) relata, também, que as tecnologias auxiliam a Web Semântica a atingir os seus objetivos, sendo estabelecidos padrões, linguagens e estruturas com o intuito de

representar, codificar, descrever, intercambiar e consultar dados, para uma futura utilização em sistemas computacionais.

No âmbito da Ciência da Informação, Santos e Alves (2009, não paginado) afirmam que as tecnologias da Web Semântica

[...] convergem para a área de Ciência da Informação, estabelecendo uma estreita relação na questão da representação do conhecimento, principalmente no que se refere ao uso de metadados considerados essenciais no estabelecimento dos requisitos para uma boa representação dos recursos informacionais na rede.

O olhar das autoras relata que a Web Semântica e suas tecnologias não estão limitadas somente à Web, ou aos objetivos da Web Semântica, verificando a importância que tais tecnologias apresentam para a Ciência da Informação, em que a implementação destas contribuiu em estudos da área.

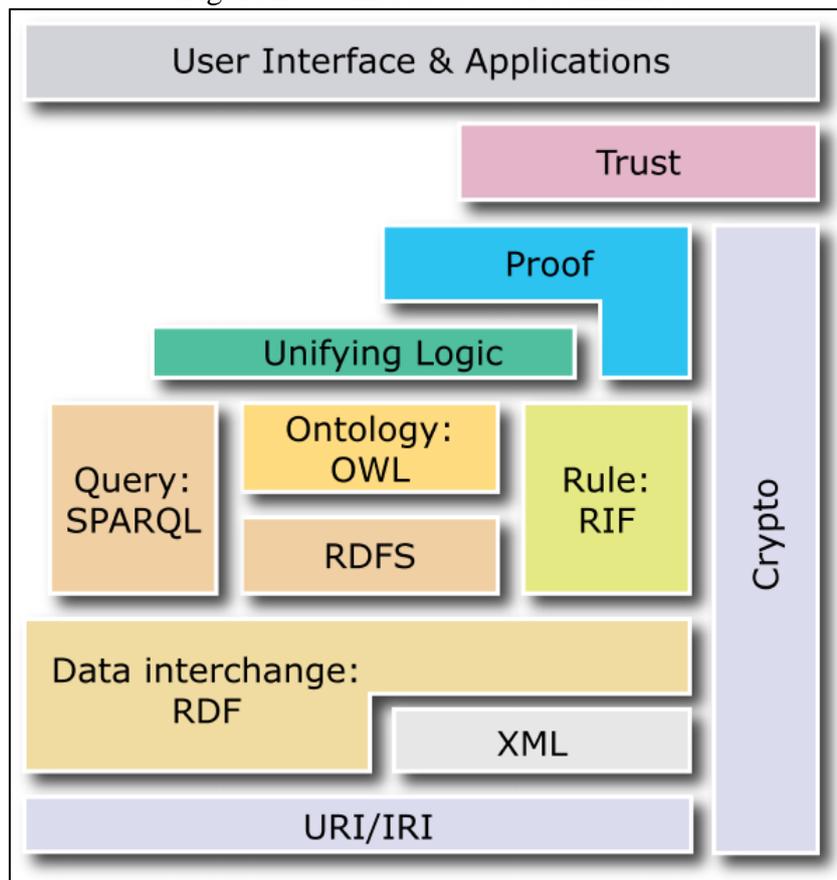
Santarem Segundo e Coneglian (2015, p. 225) reafirmam essa questão ao dizer que

Apesar da Web Semântica ser apresentada como uma estrutura de sete camadas, o passar dos anos vem nos mostrando que as tecnologias e linguagens que compõem estas camadas podem ser aplicadas individualmente ou em agrupamentos, de forma a construir ambientes digitais que permitam armazenamento e Recuperação de Informação com significado, ou seja, utilizando-se de relações semânticas.

Na explanação realizada, verifica-se que os padrões e as tecnologias da Web Semântica não se limitam apenas a Web, podendo expandir suas fronteiras para quaisquer ambientes informacionais digitais.

Para existir um melhor entendimento de como os conceitos e as tecnologias da Web Semântica funcionam e se relacionam, existe uma imagem que é chamado de Bolo de Camadas (*Layer Cake*), que exprime os diversos conceitos e as tecnologias da Web Semântica; tal imagem é representada na Figura 2.

Figura 2 - Camadas da Web Semântica



Fonte: W3C, 2007.

Na Figura 2 é possível visualizar diversos conceitos e tecnologias como URI/IRI, XML, Intercâmbio de dados (*RDF*), Consulta (*Query - SPARQL*), Ontologia (*Ontology - OWL*), Regra (*Rule - RIF*), entre outros.

Aqui, cabe uma ressalva: não é necessário que todas as camadas sejam construídas e implementadas para que existam aplicações baseadas na Web Semântica. Deve existir uma análise por parte dos desenvolvedores de um ambiente informacional digital, no sentido de se saber quais camadas deverão ser utilizadas para tornar, assim, viável a construção de aplicações baseadas na Web Semântica.

Grimaldo (2004, p.14, tradução nossa) resume parte da proposta da Web Semântica inserindo padrões e tecnologias necessárias para implementar essa proposta:

- 1- Uma linguagem que estruture os objetos digitais sintaticamente, denominado XML (*eXtensible Markup Language*);
- 2- Um formato que estruture o significado da informação que os objetos digitais possuem (em conjunto com os metadados associados), denominados RDF (*Resource Description Framework*);
- 3- Um programa de computador que recupere as informações existentes, baseado na Inteligência Artificial, denominado Agentes Inteligentes;

- 4- Um conjunto de regras que permita aos Agentes Inteligentes moverem-se dentro da Web com liberdade e de acordo com o perfil informacional do usuário, chamado de ontologia.

No contexto deste trabalho, algumas dessas camadas possuíram uma maior importância na construção do modelo proposto. Entre os conceitos e as tecnologias destacaram-se as URIs, XML, RDF, as ontologias e o OWL e o SPARQL. Tais questões serão aprofundadas a seguir, onde cada tópico será relatado detalhadamente.

3.1.1 URI

A base das camadas da Web Semântica são as URIs, que dizem respeito à identificação única de recursos na Web. Tal padrão encontra-se na base, pois identificar um recurso é fundamental para que todo o processo de semântica ocorra, além de permitir que os recursos possam ser localizados com facilidade. Assim, todos os outros conceitos terão, direta ou indiretamente, que utilizar URIs em seus processos.

A W3C (2001) define URIs como

[...] sequências curtas que identifiquem recursos na Web: documentos, imagens, arquivos para download, serviços, caixas de correio eletrônico e outros recursos. Eles fazem os recursos disponíveis sob uma variedade de esquemas de nomenclatura e métodos de acesso, tais como HTTP, FTP, e-mail e Internet endereçável da mesma maneira.

Dziekaniak e Kirinus (2004) afirmam que na linguagem humana uma palavra ou um termo pode adquirir distintos significados. As autoras relatam que tal questão traz diversos problemas em sistemas de informação, quando se busca compreender qual é o sentido utilizado de um termo dentro de um determinado contexto. A solução para isso, são URIs, pois é permitido a utilização de distintos identificadores para cada possível significado que um termo pode possuir, auxiliando para que não exista ambiguidades dentro dos sistemas.

Berners-Lee, Fielding e Masinter (2004) complementam a questão da uniformidade, dizendo que os identificadores de recursos, além de fazer com que um conceito não seja confundido a outro, possibilita também a reutilização do identificador, quando se trata do mesmo conceito, ou do mesmo objeto. Tal característica permite uma melhor recuperação de informação e melhora o acesso aos dados, pois evitará a existência de distintas descrições para um mesmo objeto. Os autores ainda relatam que as URIs apresentam; “[...] um escopo global e são interpretados consistentemente independente do contexto.” (BERNERS-LEE; FIELDING; MASINTER, 2004, p. 6).

Outro destaque na utilização de URIs é o da especificação em que uma URI pode ser uma *Uniform Resource Name* (URN) e uma *Uniform Resource Locator* (URL). A URL “[...] refere-se ao subconjunto de URIs que, além de identificar um recurso, fornece um meio de localizar o recurso através da descrição de seu mecanismo de acesso primário.” (BERNERS-LEE; FIELDING; MASINTER, 2004, p. 7). Em contrapartida, a URN refere-se ao tipo de URI “[...] que são obrigados a permanecer única e persistente globalmente, mesmo quando o recurso deixa de existir ou se torna indisponível.” (BERNERS-LEE; FIELDING; MASINTER, 2004, p. 7).

Outro conceito relacionado às URIs são os *namespaces*. Ferreira (2014) relata que os *namespaces* são um conjunto de nomes de elementos e atributos, que podem ser entendidos como vocabulários. Na prática, os *namespaces* são utilizados como prefixos, na definição de uma URI, em que já existe determinados valores, pertencentes a um vocabulário.

A identificação única dos recursos é utilizada para realizar a representação de dados, quando se utiliza as tecnologias XML, RDF e OWL. A seguir será feita uma descrição mais detalhada acerca do padrão XML.

3.1.2 XML

O uso do XML é apontado na proposta inicial da Web Semântica (BERNERS-LEE; HENDLER; LASSILA, 2001) como um meio dos usuários criarem *tags* para descreverem o conteúdo contido em uma página Web. Os autores afirmam que programas e *scripts* conseguem utilizar XML sofisticadamente, permitindo a resolução de diversas tarefas, e completam, informando que o XML não apresenta uma estrutura fixa; assim, cada usuário pode criar sua própria estrutura para descrever os conteúdos.

O padrão XML foi desenvolvido em 1996, patrocinada pelo W3C, por um grupo de trabalho chamado de *XML Working Group* (conhecido anteriormente como *SGML Editorial Review Board*), liderada por Jon Bosak, da empresa Sun Microsystems (BRAY et al., 2006).

A W3C (2015) descreve o XML como um:

[...] 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.

Bray et al. (2006) também descrevem os dez objetivos para que o XML foi criado: (i) XML deve ser diretamente utilizável na Internet; (ii) XML deve suportar uma gama ampla de aplicações; (iii) XML deve ser compatível com SGML; (iv) Deve ser fácil escrever programas que suportem XML; (v) O número de recursos opcionais do XML deve ser o mínimo possível, sendo ideal não haver nenhum; (vi) Documentos XML devem ser legíveis por pessoas, com uma clareza razoável; (vii) O projeto XML deve ser preparado rapidamente; (viii) O projeto XML deve ser formal e conciso; (ix) Deve ser fácil criar documentos XML; e (x) Concisão na marcação é de importância mínima.

A partir desses objetivos, verifica-se que o XML, em tese, deve simplificar os processos tanto para criar, quanto para interpretar, buscando ser o mais simples possível, para que haja uma grande quantidade de adeptos e para que se torne a principal linguagem de comunicação dentro da Web. Dessa forma, verifica-se que a proposta da Web Semântica tem como uma das suas camadas base o XML, por essa simplicidade, permitindo que a criação de documentos XML seja feita com facilidade e legível por máquinas com rapidez.

Decker et al. (2000) escrevem a respeito da estrutura do XML, discorrendo sobre as principais características da sintaxe da linguagem. Os autores dizem que o XML é uma linguagem de marcação para estruturar arbitrariamente um documento, em que há um conjunto organizado de *tags* de abertura e fechamento, sendo que cada *tag* pode conter inúmeros pares de informações de chave-valor, que contém características daquelas *tags*.

Um modelo para representação de dados utilizado na Web Semântica, chamado de RDF, apresenta uma maior expressividade do que o XML. Contudo, o RDF, na maioria dos casos, é construído utilizando a sintaxe XML. Na sequência, os padrões RDF e RDF *Schema* serão explicados mais detalhadamente.

3.1.3 RDF e RDF *Schema*

As primeiras aplicações com metadados dentro da Web começaram a partir de 1995, quando o W3C começou a estabelecer o *Platform for Internet Content Selection* (PICS) (MILLER, 1998). Tal especificação tinha o intuito de rotular as páginas dentro da Web, numa busca inicial de classificar os tipos de conteúdo contidos naqueles documentos. Essa especificação foi motivada inicialmente pelo intuito de criar uma classificação que pudesse informar se uma página continha conteúdos sexuais, violência, e que não fosse adequado para determinadas faixas etárias, podendo permitir que o

navegador fosse capaz de barrar os conteúdos, caso fosse configurado para isso pelos pais ou responsáveis (W3C, 2009).

Com a evolução do desenvolvimento dessa especificação, identificou-se uma necessidade de que houvesse descrições que extrapolassem as classificações restritivas de conteúdo. A partir de então, o W3C criou um grupo de trabalho chamado de *Resource Description Framework* (RDF) com a intenção de criar e discutir uma estrutura de descrição que atendesse as necessidades descritivas de diversos contextos dentro da Web (FERREIRA; SANTOS, 2013).

Uma definição que explicita os objetivos do RDF é dada por Breitman (2005, p. 20):

O Resource Description Framework (RDF) é uma linguagem declarativa que fornece uma maneira padronizada de utilizar o XML para representar metadados no formato de sentenças sobre propriedades e relacionamentos entre itens na Web. Esses itens, chamados de recursos, podem ser virtualmente qualquer objeto (texto, figura, vídeo e outros), desde que possuam um endereço Web.

O próprio W3C define o RDF como um modelo padrão para o intercâmbio de dados na Web, visto ter também o modelo RDF como característica ser estruturado para suportar a evolução dos esquemas, sem haver a necessidade de realizar correções em todos os dados, caso existam alterações (W3C, 2014).

O desenvolvimento do RDF foi motivado por cinco tópicos (W3C, 2004):

- Metadados para Web: com o objetivo de fornecer informações sobre os recursos presentes na Web.
- Aplicações que utilizam modelos de informações abertas: o RDF apresenta grande utilidade para aplicações que não contêm informações restritas, ou seja, disponibilizam seus dados para acesso aberto.
- Fazer com que as máquinas consigam processar as informações da Web: a utilização do RDF auxiliaria as máquinas a processarem os dados fora do ambiente onde foram criados, tornando possível trabalhar em grande escala na Internet.
- Interoperabilidade entre aplicações: permitir que dados de várias aplicações sejam combinados, possibilitando a inferência de novas informações.
- Processamento automatizado de informações por agentes computacionais: a Web estava em transformação, passando de um ambiente em que existiam apenas dados preparados para leitura humana, para um ambiente de

cooperação de processos, em que RDF é uma língua franca para tais processos.

Tais motivações são importantes de serem visualizadas, para que se possa compreender em quais situações o RDF foi projetado para ser utilizado, permitindo inferir em quais ambientes o uso do RDF pode agregar e permitir a aplicação da Web Semântica.

Para isso, um meio como o RDF possibilita a execução de tais tarefas perpassando, principalmente, pela sintaxe utilizada. Uma declaração (*statement*) é o conjunto de um recurso, com uma propriedade e um valor, chamados respectivamente de sujeito, predicado e objeto (W3C, 1999).

O sujeito, chamado também de recurso, representa uma informação, seja um artigo científico, uma música, um autor, entre outros, sendo qualquer coisa que possua uma URI.

O predicado, ou propriedade, 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 dizendo que “[...] são os atributos que permitem distinguir um recurso de outro ou que descrevem o relacionamento entre recursos”.

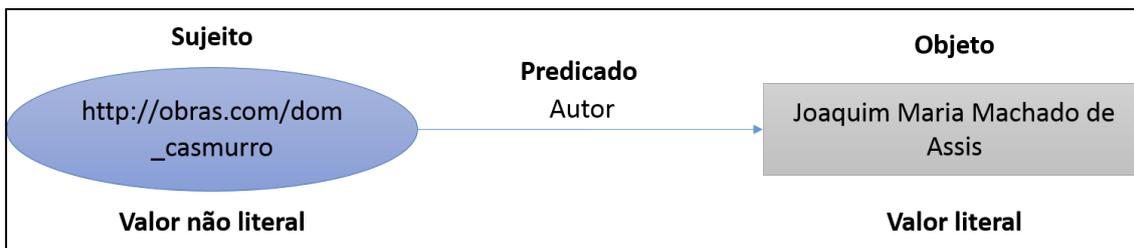
Por fim, o objeto, ou valor, são os dados que representam o conteúdo que está sendo descrito. Breitman (2005) diz, ainda, que os objetos são os únicos elementos do RDF que podem ser tanto um recurso, quanto um valor literal.

Para compreender o significado de um valor literal, torna-se necessário diferenciar valores literais de valores não literais. Os valores não literais são os recursos em si, sendo utilizados quando os valores podem ser representados por meio de URIs. Contudo, os valores literais são utilizados quando se lida com valores de dados concretos, como números inteiros e reais, valores absolutos e cadeias de caracteres, chamadas de *string* (FERREIRA; SANTOS, 2013).

A Figura 3 demonstra, por meio de representação gráfica do RDF, os diversos elementos explicitados no texto. O primeiro elemento destacado são os arcos, que sempre farão referência a um sujeito e contém uma URI. Outro ponto de destaque são os valores literais e não literais; os valores literais são representados em retângulos que, como dito, contém uma *string*, um número ou um valor absoluto, no entanto os valores não literais, como representam recursos, são indicados pelos arcos.

A Figura 3 contém uma declaração, em que o sujeito é indicado pelo recurso com a URI “http://obras.com/dom_casmurro”, o predicado é a propriedade “autor” e o objeto é um valor literal, com o texto “Joaquim Maria Machado de Assis”.

Figura 3 – Exemplo - representação gráfica do RDF



Fonte: Elaborado pelo autor

No momento, a principal forma de codificar o RDF é por meio do XML, chamado de RDF/XML. Porém, existem diversas outras sintaxes que podem ser utilizadas, como JSON, Turtle e N-Triples (FERREIRA; SANTOS, 2013). A Figura 4 representa o conteúdo apresentado na Figura 3, utilizando o XML para realizar a descrição.

Figura 4 - Representação RDF/XML

```

1: <?xml version="1.0" encoding="ISO-8859-1" ?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
4:
5: <rdf:Description rdf:about="http://obras.com/dom_casmurro">
6:
7:   <autor>
8:     <rdf:Description rdf:about="http://autores.com/machado_de_assis">
9:       <nome>Joaquim Maria Machado de Assis</nome>
10:    </rdf:Description>
11:  </autor>
12:
13: </rdf:Description>

```

Fonte: Elaborado pelo autor.

Santarem Segundo (2010) explicita uma limitação existente na utilização do RDF, pois esse padrão contém um número reduzido de elementos pré-definidos, tornando inviável o desenvolvimento de vocabulários próprios por comunidades independentes, não permitindo que ontologias sejam construídas utilizando o RDF. O autor complementa dizendo que o RDF apresenta diversos campos que podem ser aplicados, em que as necessidades se concentram na realização de relacionamentos em si. Entretanto, para realizar a definição de dados, pelo fato do RDF não se mostrar adequado, a W3C desenvolveu o *RDF Schema*.

Breslin, Passant e Decker (2009) dizem que o propósito do *RDF Schema* é definir classes, instâncias e relacionamentos, em que o *RDF Schema* seria uma aplicação do

RDF. Breitman (2005) completa afirmando que o RDF *Schema* não fornece as classes e as propriedades em si, mas, sim, é um *framework* que possibilita a descrição.

A W3C (2014b) define o RDF *Schema* como uma extensão semântica do RDF, que provê um vocabulário de modelagem de dados para RDF. É dito, ainda, que esse padrão provê mecanismos para realizar a descrição de grupos de recursos, bem como dos relacionamentos existentes entre estes.

As principais classes que compõem o RDF *Schema* são descritas no Quadro 2, que contém os nomes das classes e uma breve descrição do significado de cada classe.

Quadro 2 - Principais classes RDF *Schema*

Classe	Descrição
<i>rdfs:Resource</i>	A classe de todos os recursos.
<i>rdfs:Class</i>	A classe que todas as classes RDF pertencem.
<i>rdfs:Literal</i>	A classe de todos os literais, como <i>strings</i> e números.
<i>rdfs:Datatype</i>	A classe dos tipos de dados.
<i>rdf:langString</i>	A classe dos valores descritos em <i>strings</i> .
<i>rdf:Html</i>	A classe de valores HTML.
<i>rdf:XMLLiteral</i>	A classe de valores XML.
<i>rdf:Property</i>	A classe das propriedades RDF.

Fonte: Adaptado (W3C, 2014; BREITMAN, 2005, p. 52)

No Quadro 2, verifica-se que algumas das classes são elementos do RDF, como o “*rdf:langString*”. Esse fato ocorre, pois como dito, o RDF *Schema* é uma extensão do RDF, sendo utilizado diversos elementos do RDF nesse padrão. Para identificar se uma classe é oriunda do RDF ou do RDF *Schema*, deve-se verificar o início das notações das classes, em que as classes iniciadas com “*rdf:*” pertence ao RDF e as iniciadas com “*rdfs:*” são pertencentes ao RDF *Schema*. Outra característica relevante do quadro, é que existem elementos que são especificações de outros, como no caso do “*rdf:langString*” que é uma subclasse do “*rdfs:literal*”, e é uma instância da classe “*rdfs:Datatype*”. Por isso, diversas classes são enquadradas em vários tipos de elementos.

Outro elemento do *RDF Schema* são as propriedades, que possibilitam a realização de relacionamentos, inserindo semântica na descrição realizada dos dados. O Quadro 3 representa todas as propriedades descritas pela W3C que o *RDF Schema* possui.

Quadro 3 - Propriedades *RDF Schema*

Propriedade	Descrição
<i>rdfs:range</i>	Tem a função de indicar que os valores de uma propriedade são instâncias de uma ou mais classes.
<i>rdfs:domain</i>	Tem a função de indicar que qualquer recurso que possui uma determinada propriedade é uma instância de uma ou mais classes.
<i>rdf:type</i>	É utilizado para indicar que um recurso é uma instância de uma determinada classe.
<i>rdfs:subClassOf</i>	É utilizado para indicar que todas as instâncias de uma classe são instâncias de uma outra.
<i>rdfs:subPropertyOf</i>	É utilizado para indicar que todos os recursos relacionados a uma propriedade estão relacionados a uma outra.
<i>rdfs:label</i>	Provê informações para a leitura humana do nome de um recurso.
<i>rdfs:coment</i>	Provê informações para a leitura humana da descrição de um recurso.

Fonte: Adaptado (W3C, 2014)

Uma informação relevante das propriedades descritas no Quadro 3, é que todas elas são instâncias de “*rdf:Property*”, que é uma classe padrão das propriedades pertencentes ao RDF. Outro destaque, são as classes *range* e *domain*, que significam respectivamente alcance e domínio, tendo como função ligar duas instâncias por meio de uma propriedade, pois para que dois objetos tenham relação, sempre existirá uma instância (*domain*) que possui um valor de uma outra instância (*range*).

Com as informações relatadas, é possível verificar uma porção das possibilidades que tanto o RDF quanto *RDF Schema* permitem. Diversos autores destacam, ainda, que o *RDF Schema* pode ser utilizado como uma linguagem para a construção de ontologias. Para compreender melhor as ontologias, a próxima subseção trata desse conceito e da principal linguagem utilizada na construção de ontologias, que é a *Web Ontology Language* (OWL).

3.1.4 Ontologias

O termo ontologia é derivado do idioma grego, *onto* (ser) e *logia* (estudo), tendo seus estudos iniciais dentro da Filosofia. Ramalho (2006) diz ainda que apesar de ser utilizado desde os tempos de Platão e Aristóteles, o termo ontologia fazendo referência a um ramo da Filosofia surge em uma época muito mais contemporânea, em meados dos séculos XVII e XVIII.

Mais recentemente, o termo ontologia começou a ser utilizado em pesquisas de Ciência da Computação e Ciência da Informação. Com um foco diferente, estando mais relacionado a descrever o contexto de um domínio existente no mundo, de modo que ficasse computacionalmente clara uma situação do mundo real.

Ramalho (2006) complementa dizendo que as pesquisas com maior impacto começaram a ser desenvolvidas dentro da Ciência da Computação na década de 1990, principalmente com foco em organizações de bases de conhecimento, dentro da disciplina de Inteligência Artificial.

Uma definição bastante utilizada na literatura, para definir ontologias, é dada por Gruber (1993, p. 1, tradução nossa), considerando que “uma ontologia é uma especificação explícita de uma conceitualização”, em que especificação explícita é compreendida como uma representação de conceitos e relacionamentos expressos por meio de símbolos e conceitualização, por se tratar de um modelo abstrato de algum domínio.

Posteriormente Borst (1997, p. 12, tradução nossa) complementa a definição dada por Gruber, relatando que: “[...] uma ontologia é uma especificação formal de uma conceitualização compartilhada”. O autor insere dois elementos, formal e compartilhada. Entende-se formal como uma representação que possa ser processável e entendida por computadores, e compartilhada por representar um conhecimento consensual.

Guarino apresenta outra definição bastante utilizada na literatura, que se relaciona fortemente com a definição dada por Borst. Guarino (1998, p. 5, tradução nossa) afirma que “Uma ontologia é uma teoria lógica que representa o significado pretendido de um vocabulário formal”.

Ramalho (2006, p. 57) relaciona as definições dada por Guarino (1998) e Gruber (1993), explicando 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.

Uma definição mais recente é dada por Santarem Segundo (2010), embasada em Gruber (1993), Borst (1997) e Guarino (1998), no entanto apresentando uma visão mais relacionada com a área da Ciência da Informação. O autor expõe que as ontologias são construídas pela

[...] necessidade de um vocabulário compartilhado em que um conjunto de informações possam ser trocadas e também reusadas pelos usuários de uma comunidade. Considere os usuários de uma comunidade seres humanos ou agentes inteligentes (SANTAREM SEGUNDO, 2010, p. 104)

Nessa definição, verifica-se a inserção de elementos fundamentais quando se trata de ontologias, tal como a questão dos vocabulários compartilhados, sendo a ontologia uma contextualização de um domínio, que pode ser considerado um vocabulário, que poderá ser reusado para diversas situações e usuários.

Uma definição que insere ainda mais profundamente elementos da Ciência da Informação é dada por Campos e Campos (2014, p. 3827-3828). Elas afirmam 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, ao inserirem questões como os pressupostos epistemológicos e ontológicos, incluem o fato de que as ontologias devem ser uma representação bastante fiel, que possibilita a expressão de uma área do conhecimento e de um domínio, por meio do uso de ontologias. Por tais motivos, elas dizem que o processo de elaboração de ontologias é bastante custoso.

Por meio de todas essas definições, verificam-se características fundamentais dentro do conceito de ontologias, em que as ontologias apresentam um modelo informacional computacional que representa determinados contextos, sendo que tais modelos devem buscar ser consensuais entre as comunidades relacionadas àquele determinado domínio. As definições de Santarem Segundo (2010) e Campos e Campos (2014) permitem inferir que as ontologias podem ser utilizadas como vocabulários compartilhados, para diversos domínios, especialmente aqueles relacionados a iniciativas de dados abertos, que devem necessariamente possuir vocabulários compartilhados.

Outra característica fundamental, no tratamento de ontologias, diz respeito às classificações que estas podem possuir conforme o objetivo de uso que possuem. Guarino (1998, p.10-11) descreve quatro tipos principais de ontologias:

- Ontologias de topo (*top-level ontologies*): têm a função de descrever conceitos bastante gerais, como espaço, tempo, matéria, objeto, evento, ação, entre outros. As representações feitas por ontologias desse tipo, não dependem de um domínio ou de um problema específico.
- Ontologias de domínio (*domain ontologies*): têm a função de descrever um domínio particular, “especializando conceitos introduzidos nas ontologias de topo” (SANTAREM SEGUNDO, 2010, p. 104). Alguns exemplos seriam ontologias tratando de medicina e automobilismo.
- Ontologias de tarefa (*task ontologies*): têm a função de descrever uma tarefa ou atividade genérica, dentro de um domínio. Exemplos seriam os prontuários médicos dentro do domínio de medicina ou a compra e venda de veículos, no domínio automobilístico.
- Ontologias de aplicação (*application ontologies*): têm a função de descrever uma aplicação, que depende tanto de um domínio quanto de uma tarefa, e por vezes é uma especificação de ontologias de ambos os tipos (ontologias de domínio e de tarefa). Tais conceitos, diversas vezes, correspondem ao papel desempenhado por algum ator dentro de um domínio, durante a execução de uma determinada tarefa.

A classificação proposta por Guarino tem importância, no momento de avaliar qual será o tipo de ontologia que deverá ser utilizada dentro de um contexto, seja para construir uma nova ontologia, seja para buscar por ontologias que já foram elaboradas. Outro ponto diz respeito ao reaproveitamento de ontologias; portanto, estando claro o tipo de ontologias que se deseja construir, pode-se utilizar uma outra ontologia de base, seja ela de topo ou de domínio.

A relação entre ontologias e Web Semântica começou a ser desenhada na proposta da Web Semântica no texto de Berners-Lee, Hendler e Lassila (2001). Nesse texto, os autores relatam a necessidade da Web Semântica possuir vocabulários que permitam ser realizadas inferências, em que estejam definidas regras, para conduzir os agentes computacionais a processarem a semântica do contexto de cada página.

Santarem Segundo e Coneglian (2015, p. 227) complementam o conceito de ontologias, no contexto das aplicações da Web Semântica, dizendo que “[...] entende-se

as ontologias como: artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de: classes, propriedades, relações, restrições, axiomas e instâncias”.

Por meio dessas definições, é possível verificar a importância do papel das ontologias, na Web Semântica, e quanto elas são necessárias, quando se trata de explicitar as relações existentes entre os dados, com base no domínio ao qual pertencem as informações. Resumindo, a ontologia, no contexto da Web Semântica, apresenta diversas funções, como a de evitar a ambiguidade de sentidos dos dados, realizar inferências, e permitir que diversos tipos de relações ocorram.

Catarino e Souza (2012) descrevem que a Web Semântica necessita ter vocabulários para representarem conceitos, relacionamentos, restrições e propriedades entre os dados. As autoras descrevem ainda que não existe uma recomendação da W3C, de qual tecnologia utilizar, mas elas relatam que se costuma utilizar vocabulários mais simples para realizar a descrição de coleções mais simples, e o uso de ontologias para coleções mais complexas.

No entanto, como as ontologias apresentam grande possibilidade de descrição, e permitem uma grande variedade de opções de contextualizar um domínio, inserindo uma semântica mais consistente, tal conceito aparece como opção mais viável e eficiente na construção desses vocabulários.

Para compreender o conceito de ontologia utilizada neste trabalho, torna-se necessário discorrer acerca dos diversos tipos de vocabulários existentes. Breitman (2005) descreve diversos tipos de vocabulários que podem ser utilizadas na Web Semântica e também em diversos outros contextos. O relato de Breitman (2005) baseia-se na classificação realizada por Lassila e McGuinness (2001). Estas autoras realizaram uma categorização dos vocabulários, na seguinte sequência: 1) Vocabulários controlados/catálogos; 2) Termos/Glossário; 3) Tesouros; 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.

Cabe ressaltar a distinção entre dois conceitos principais, taxonomias e tesouros, e ressaltar que, por vezes, ontologias são confundidas com eles. Breitman (2005, p. 34) diz que “[...] uma taxonomia serve para classificar uma informação em uma hierarquia (árvore), utilizando o relacionamento pai-filho (generalização ou tipo-de).” A mesma autora (BREITMAN, 2005, p. 36) define tesouro como “[...] uma taxonomia adicionada

de um conjunto de relacionamentos semânticos (equivalência, associação, entre outros) entre seus termos.”

Partindo dessa definição, verifica-se que os conceitos de ontologias apresentados demonstram diferenças substanciais, quando comparado a taxonomias e tesouros. As ontologias, além de apresentar relações hierárquicas e os relacionamentos semânticos que um tesouro possui, dispõem de restrições de valores e permitem a inserção de lógica na definição da semântica, sendo esta a concepção de ontologia utilizada no âmbito desta pesquisa. Na classificação feita por Lassila e McGuinness (2001), as categorias 6, 7 e 8 apresentam as características que as diferenciam de tesouros e taxonomias.

No âmbito deste trabalho, ontologias serão tratadas como diferentes de tesouros e taxonomias, utilizando essencialmente as características que as distinguem, para permitir a inserção de semântica na proposta que será especificada adiante.

Uma tecnologia de fundamental importância neste trabalho é a principal linguagem de implementação de ontologias utilizada atualmente, a OWL. Cabe ressaltar que o termo ontologia refere-se a um conceito, e há necessidade de que haja uma linguagem que implemente as características das ontologias; nesse contexto surge a linguagem OWL. A seguir essa linguagem será tratada com mais detalhes.

3.1.4.1 OWL

A linguagem OWL foi desenvolvida pela W3C como uma iniciativa de permitir uma melhor representação, mais expressiva e mais complexa, do que aquela possibilitada pelas tecnologias RDF e RDF *Schema*. Tal tecnologia é uma revisão da linguagem DAML+OIL, que possibilitava a representação de ontologias, porém não apresentava uma capacidade representacional tão extensa quanto a OWL (BRESLIN; PASSANT; DECKER, 2009).

O grupo de trabalho da OWL, da W3C, define tal linguagem como:

[..] 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 (W3C, 2012, tradução nossa).

Santarem Segundo (2010) complementa dizendo que OWL é a principal linguagem existente para a construção de ontologias, e tem o intuito de atender às necessidades de aplicação da Web Semântica. O autor completa relatando os objetivos para o desenvolvimento da OWL:

[...] construir ontologias, 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 (SANTAREM SEGUNDO, 2010, p. 128).

Breitman (2005) relata a existência de seis elementos básicos na OWL, que são: *namespaces*, cabeçalhos, classes, indivíduos, propriedades, e restrições de classes e propriedades. É a partir do uso de tais elementos que a descrição de um domínio pode ser realizada, permitindo apresentar uma semântica explícita dos dados.

O primeiro componente relatado é o *namespace*, que possui uma relação com as URIs, tendo a função de inserir um conjunto de elementos que podem ser encontrados por meio de uma localização única, evitando, dessa forma, ambiguidades. Normalmente, uma ontologia construída em OWL sempre inicia com declarações de um conjunto de *namespaces*, que permitem definir um *namespace* para os próprios elementos da ontologia, e a definição de *namespaces* de elementos externos que são utilizadas na mesma ontologia. Um exemplo de *namespace* é o “*dcterms*”², que representa os termos do padrão de metadados do *Dublin Core*.

Na estrutura de uma ontologia, os cabeçalhos aparecem posteriormente a definição dos *namespaces*, sendo elementos utilizados para a definição de sentenças, tratando da própria ontologia. Basicamente, insere, nos cabeçalhos, os comentários, o controle das versões e a inclusão de conceitos e propriedades de outras ontologias.

As classes apresentam-se como um dos elementos essenciais das ontologias, pois são elas que irão representar os objetos do mundo real descritos em uma ontologia. Breitman (2005, p. 61) diz 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 [...]”. Santarem Segundo (2010, p. 133) complementa dizendo que “[...] a classe é utilizada para definir o conceito abstrato de um determinado

² DCMI Metadata Terms. Disponível em: <<http://dublincore.org/documents/dcmi-terms/>>. Acesso em: 10 mai. 2016.

domínio [...]”. Bechhofer et al. (2004) diz ainda que “[...] a classe define um grupo de indivíduos que pertencem a tal classe, pois compartilham algumas propriedades”.

Tais definições explicitam que a classe é uma abstração de algo que será representado em uma ontologia, e o objeto concreto de tais classes são os indivíduos, sendo que um indivíduo pertencente a uma classe carrega todas as características, propriedades e relações dela.

Todas as classes definidas em OWL são subclasses dessa superclasse chamada “owl:Thing”, o que significa, realizando uma abstração, que todas as definições que essas classes estão representando são inevitavelmente uma coisa, seja ela pessoa, objeto, conceito, entre outras. Outra característica das classes OWL, são que elas são uma instância nomeada da classe “owl:class”, que por sua vez é uma subclasse de “rdfs:Class” (BECHHOFER et al., 2004).

A linguagem OWL possibilita a definição de subclasses. As subclasses representam uma especificação de uma superclasse, ou seja, uma classe apresenta determinadas características, e uma subclasse desta irá apresentar as mesmas características, acrescidas de outras distintas. A definição de uma subclasse ocorre por meio da utilização do termo “rdfs:subClassOf”, que irá definir o relacionamento entre as classes.

Outro elemento das ontologias OWL, são os indivíduos. Breitman (2005, p. 62) diz que “[...] indivíduos são objetos do mundo; pertencem a classes e são relacionados a outros indivíduos (e classes) através de propriedades. Indivíduos são os membros das classes”. Bechhofer et al. (2004) dizem que “os indivíduos são instâncias de classes, e as propriedades podem ser usadas para relacionar um indivíduo a outro”.

O quinto elemento da OWL são as propriedades. As propriedades são definições que caracterizam as classes e, por sua vez, os indivíduos. Santarem Segundo (2010, p. 135) diz 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”.

Antoniou e Harmelen (2004) discorrem a respeito dos tipos de propriedades que a linguagem OWL possui: as propriedades de objeto (*object properties*), que relacionam objetos entre si, e as propriedades de dados, que relacionam um objeto com dados valorados. Um exemplo de propriedades de objeto seria a propriedade “trabalha em”, relacionando uma pessoa a uma empresa. Já um exemplo de propriedades de dados seria “idade”, que pode relacionar uma pessoa a um número inteiro.

O último elemento básico da OWL são as restrições, que são aplicadas nas propriedades, e como o próprio nome sugere, têm a função de restringir os valores que podem ser aplicados às propriedades. Santarem Segundo (2010) classifica em dois os tipos de restrições, as de cardinalidade e as de valores.

As restrições de cardinalidade buscam definir o número arbitrário de valores que uma instância de uma classe pode ter para uma determinada propriedade, podendo ser aplicadas tanto para Propriedades de Dados, quanto para Propriedades de Objetos. 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 a 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 (W3C, 2012b) insere ainda a possibilidade de inserção de cardinalidade qualificada, por meio das propriedades “owl:maxQualifiedCardinality”, “owl:minQualifiedCardinality” e “owl:qualifiedCardinality”, que apresenta as mesmas características que “owl:maxCardinality”, “owl:minCardinality” e “owl:cardinality”, respectivamente. No entanto, quando é cardinalidade qualificada, torna-se obrigatório especificar, além da propriedade, a qual classe está sendo especificada aquela cardinalidade.

As restrições de valores são divididas em três tipos: “owl:allValuesFrom”, “owl:someValuesFrom” e “hasValue”.

- Todos os valores de (owl:allValuesFrom): tem a função de definir quais são os valores possíveis que uma propriedade determinada pode ter.
- Tem o valor (owl:hasValue): tem a função de definir um valor determinado que uma determinada propriedade especificada pode possuir.
- Algum valor de (owl:someValuesFrom): tem a função de determinar a classe e a ocorrência de pelo menos um valor dentre as propriedades.

Com o lançamento da segunda versão da OWL (W3C, 2012b) foi inserido outro elemento para restrições de valores:

- Tem ele mesmo (owl:hasSelf): tem a função de especificar uma ligação de um indivíduo nele mesmo.

Dentro desse grupo de restrições, há algumas propriedades especiais que permitem a criação e exploração de axiomas. Tais propriedades permitem a definição de diversas características e relacionamentos entre propriedades e classes. Parte delas foram desenvolvidas na primeira versão do OWL, e outras na segunda versão.

O termo propriedades especiais é utilizado por Antoniou e Harmelen (2004); em contrapartida, a W3C na primeira versão do OWL chama tais elementos de características de propriedades, já na segunda versão existe uma maior separação entre os elementos, inserindo diversos novos elementos, tendo como nomes: axiomas de expressão de classes, axiomas de propriedades de objetos, axiomas de propriedades de dados e afirmações, conectivos booleanos e enumeração de indivíduos (W3C, 2012). O Quadro 4 foi construído com a intenção de reunir as principais propriedades especiais, que apresentam a função de inserir uma semântica ainda mais elaborada e forte nas ontologias construídas em OWL. Nesse quadro não foram inseridos os elementos de restrições de propriedades, pois elas já foram enumeradas anteriormente; tampouco foram consideradas as propriedades relacionadas a declarações de classes e propriedades, pois estão relacionadas ao momento de criação das ontologias e não ao processo de descoberta e visualização da semântica contida em um domínio.

Quadro 4 - Propriedades Especiais OWL

Nome	Notação	Classificação	Característica
Classes equivalentes	<i>owl:equivalentClass</i>	Axiomas – Classes	Essa propriedade tem a função de representar que duas classes apresentam semântica equivalente.
Classes disjuntas	<i>owl:disjointWith</i>	Axiomas – Classes	Um indivíduo pode ser instância de várias classes; entretanto, quando duas classes são disjuntas, um indivíduo não pode ser instância dessas duas classes.
Classes disjuntas aos pares	<i>owl:AllDisjointClasses</i>	Axiomas – Classes	Permite a disjunção de diversas classes disjuntas, não somente de duas.
União de disjunção	<i>owl:disjointUnion Of</i>	Axiomas – Classes	Permite a disjunção de uma classe para diversas outras classes.

Nome	Notação	Classificação	Característica
Inclusão cadeia de propriedades	<i>owl:propertyChainAxiom</i>	Axiomas – Propriedades	Permite definir um relacionamento a partir da definição de outras duas definições de relacionamentos
Propriedades equivalentes	<i>owl:equivalentProperty</i>	Axiomas – Propriedades	Define que duas propriedades distintas apresentam equivalência na semântica apresentada.
Disjunção de propriedades	<i>owl:propertyDisjointWith</i>	Axiomas – Propriedades	Semelhante à disjunção de classes, pois, um indivíduo pode ser instância de duas propriedades, entretanto, se duas propriedades forem disjuntas, um indivíduo não pode ser instância de ambas.
Disjunção de propriedades aos pares	<i>owl:AllDisjointProperties</i>	Axiomas – Propriedades	Permite a disjunção entre diversas propriedades.
Propriedades inversas	<i>owl:inverseOf</i>	Axiomas – Propriedades	Uma propriedade por ser especificada como sendo o inverso de uma segunda.
Propriedades funcionais	<i>owl:FunctionalProperty</i>	Axiomas – Propriedades	Define que uma propriedade desse tipo deve ter um único valor; sendo utilizada diversas vezes como um atalho para indicar que a cardinalidade mínima é 0 e a máxima é 1.
Propriedades funcionais inversas	<i>owl:InverseFunctionalProperty</i>	Axiomas – Propriedades	Em um relacionamento de propriedades de objetos, é possível definir qual relacionamento é o inverso de uma propriedade funcional.
Propriedade reflexiva	<i>owl:ReflexiveProperty</i>	Axiomas – Propriedades	Define que uma propriedade é reflexiva, quando uma classe se relaciona a si mesma, por meio dessa propriedade.
Propriedade irreflexiva	<i>owl:IrreflexiveProperty</i>	Axiomas – Propriedades	Uma propriedade irreflexiva ocorre quando um tipo de relacionamento não pode ocorrer em uma própria classe. Por exemplo, ninguém pode ser pai de si mesmo, por tanto o relacionamento “é_pai_de” é irreflexivo.

Nome	Notação	Classificação	Característica
Propriedade simétrica	<i>owl:Symmetric Property</i>	Axiomas – Propriedades	Uma propriedade simétrica indica que se há um relacionamento X, de A para B, há também um relacionamento X, de B para A.
Propriedade assimétrica	<i>owl:Asymmetric Property</i>	Axiomas – Propriedades	Ocorre quando uma propriedade não é simétrica, ou seja, quando um relacionamento X, de A para B, não vale de B para A.
Propriedade transitiva	<i>owl:Transitive Property</i>	Axiomas – Propriedades	Uma propriedade pode possuir características transitivas, ou seja, se A possui relação X com B, e se B possui relação X com C, A possui relação X com C.
Igualdade de indivíduos	<i>owl:sameAs</i>	Axiomas - Afirmações	Define que diversos indivíduos diferentes são semelhantes
Diferença de indivíduos	<i>owl:differentFrom</i>	Axiomas - Afirmações	Define que dois indivíduos não se referem a um único indivíduo.
Diferença de indivíduos aos pares	<i>owl:AllDifferent</i>	Axiomas - Afirmações	Define que diversos indivíduos não se referem a um mesmo indivíduo.
Afirmação de propriedades negativas	<i>owl:Negative PropertyAssertion</i>	Axiomas - Afirmações	Essa propriedade define que dois indivíduos não estão relacionados por meio de uma relação.
Intersecção	<i>owl:intersectionOf</i>	Conectivos Booleano	Uma intersecção entre classes aponta os indivíduos que são instâncias das duas classes.
União	<i>owl:unionOf</i>	Conectivos Booleano	Uma união entre classes aponta todos os indivíduos que são instâncias de pelo menos uma das classes.
Complemento	<i>owl:complementOf</i>	Conectivos Booleano	O complemento aponta os indivíduos que não são membros de uma determinada classe.
Enumeração	<i>owl:oneOf</i>	Conectivos Booleano	Descreve todas as instâncias de uma classe por meio de uma enumeração.

Fonte: Elaborado pelo autor.

Por meio do Quadro 4, é possível visualizar a grande quantidade de propriedades que a OWL possui para realizar a descrição do contexto de um domínio, possibilitando a inserção de inúmeras características, que conseguem representar todos os elementos contidos no domínio, além de representar os relacionamentos existentes.

As características da linguagem OWL têm mudado a concepção das ontologias, e expandido sua capacidade interpretativa, descritiva e semântica de apresentar contextos.

Para manipular dados de ontologias construídas em OWL, e dados representados em RDF e RDF *Schema*, foi desenvolvido a linguagem SPARQL. Na sequência será abordada essa linguagem de consultas em dados estruturados.

3.1.5 SPARQL

A linguagem SPARQL é uma importante ferramenta dentro da Web Semântica, uma vez que é por meio dessa linguagem de consultas que se torna possível realizar buscas em dados estruturados em RDF e OWL. A primeira versão dessa linguagem de consulta foi lançada em janeiro de 2008 (W3C, 2008) e, posteriormente, em 2013 foi lançada a versão 1.1 (W3C, 2013), que apresenta diversas mudanças, sendo uma tecnologia com maior robustez e mais maturidade.

Santarem Segundo (2014, p. 3870) afirma que “SPARQL é um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web.” O autor complementa relatando que o SPARQL possibilita a utilização dos dados da Web Semântica como um todo. Tal afirmação é feita, pois implementações baseadas nas tecnologias da Web Semântica apresentam a característica de armazenar dados com formato estruturado em *data sets* (bases de dados), necessitando possuir uma linguagem que consiga extrair os dados dessas bases de dados.

Esse argumento dado pelo autor pode ser complementado pela necessidade de contextualização que os dados estruturados podem ter. Nesse âmbito, representações feitas em OWL e RDF podem possuir propriedades que permitam a execução de inferências e lógicas que estão explicitadas nas representações, que somente uma linguagem de consultas construída com base nessas tecnologias seria capaz de identificar e executar. Dessa forma, uma extração utilizando tecnologias tradicionais de banco de dados, como SQL, não permitiria uma obtenção de todas as características que tais representações possuem.

Para compreender melhor a importância do SPARQL para a Web Semântica, Berners-Lee (W3C, 2008b, tradução nossa) 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.”

Breslin, Passant e Decker (2009) complementam a definição do SPARQL, ao relatar que, como os dados em RDF são representados por meio de grafos, o SPARQL é, assim, uma linguagem de consultas por grafos, diferenciando-se do SQL, pois esta linguagem realiza consultas unicamente por meio de tabelas e de colunas.

O intuito principal do SPARQL é realizar a recuperação de dados em formato RDF, consultando os dados que estão nativamente armazenados nesse padrão, ou que podem nele ser transformados, por algum serviço. Cabe ressaltar que a linguagem OWL é expressa por meio do RDF, sendo uma camada de nível superior, que utiliza as estruturas do RDF; dessa forma, como o SPARQL realiza buscas em RDF, implicitamente, fica claro que é possível recuperar dados representados em OWL. O retorno de uma consulta construída em SPARQL são conjuntos de dados ou grafos RDF (W3C, 2013).

Além disso, o SPARQL permite a realização de inferências a respeito dos dados, pois é possível fazer diversos tipos de operações com as triplas RDF, tais como agregação, consulta dentro de uma outra consulta, negação, expressões mais complexas, entre outras (DUCHARME, 2011).

Ducharme (2011) afirma que existem quatro formas de realizar uma consulta em SPARQL:

- *SELECT*: forma principal de realizar consultas dentro do SPARQL, em que são recuperadas informações seguindo uma sintaxe e um padrão definido.
- *CONSTRUCT*: tem a função de retornar triplas RDF. É utilizado quando há a necessidade de se construir novos dados, a partir de combinações, conversões e replicações de dados anteriores.
- *ASK*: o processador de consultas verifica se um determinado padrão de grafos descreve um conjunto de triplas RDF de uma base de dados específica, retornando operadores booleanos, de verdadeiro ou falso.
- *DESCRIBE*: tem a função de identificar e descrever todas as triplas relacionadas a um objeto particular, especificado na consulta.

Esses quatro modos de consultar dados apresentados pelo SPARQL, demonstra que tal linguagem permite uma grande quantidade de funções, permitindo explorar diversas questões, conforme as necessidades dos usuários.

No entanto, a principal forma de consultar os dados com o SPARQL, se dá por meio do *SELECT*. O *SELECT* dentro do SPARQL contém quatro partes essenciais: *prefix*, *select*, *from* e *where*. Existem também os modificadores, como o *order by*, porém apresenta uma função auxiliar às quatro partes essenciais que foram destacadas. A Figura 5 mostra um exemplo de uma consulta SPARQL.

Figura 5 - Exemplo de uma consulta SPARQL

```

PREFIX p: <http://dissertacao.com/ppgci/pessoa#>
PREFIX b: <http://dissertacao.com/ppgci/banco#>
SELECT ?cpf ?nome ?saldo
FROM <banco.ttl>
WHERE
{
?pessoa      p:nome      ?nome ;
              p:cpf        ?cpf .
?banco       p:cpf        ?cpf ;
              b:saldo      ?saldo .
}

```

Fonte: Elaborado pelo autor.

Na Figura 5 é possível visualizar as quatro partes supracitadas na sintaxe do *SELECT*, na ordem em que elas devem ser inseridas, para ocorrer um funcionamento correto na consulta.

O primeiro elemento, o *prefix*, tem a função de definir os *namespaces* que serão utilizados nas consultas. A definição dos *namespaces* faz com que os resultados obtidos pertencentes a tais *namespaces*, apareçam com o prefixo definido.

O *select* denota função relacionada à apresentação dos dados. É nesse fragmento que deverão ser definidas as variáveis que deverão ser mostradas ao usuário. Além disso, para agregar mais significado, é possível realizar a apresentação dos dados, com diversas modificações, por meio de operações matemáticas, ou com um agrupamento de valores, permitindo demonstrar informações como soma, contagem e média de alguns dados.

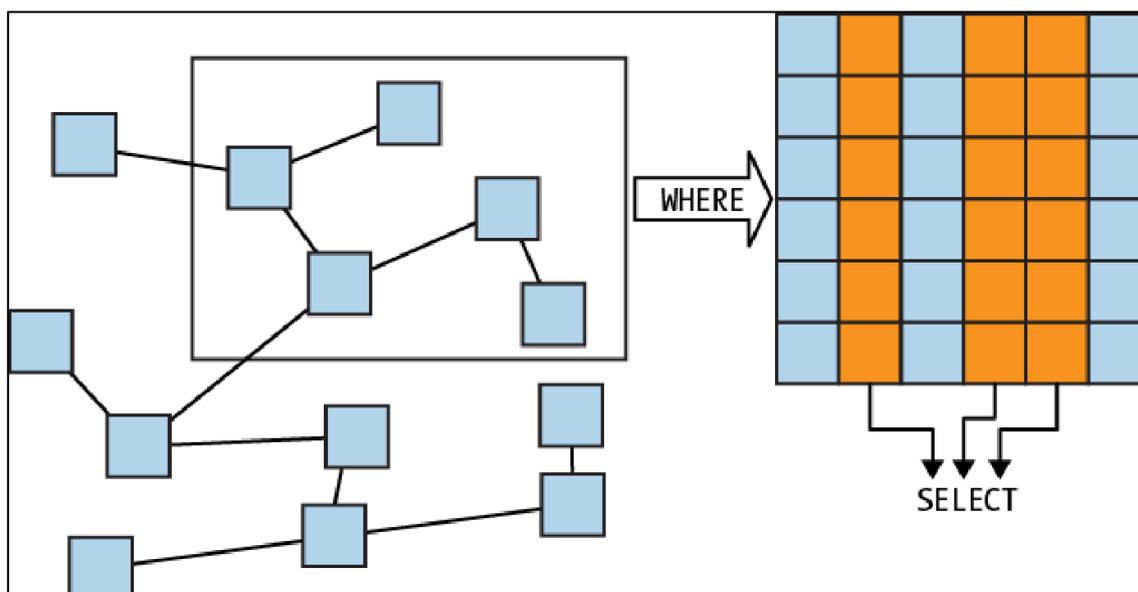
A cláusula *from* indica qual será a fonte de dados utilizada para a realização da consulta, sendo inseridos nesse local, tanto endereços que apontam para arquivos de dados internos, quanto endereços de bases de dados externas. Tal fragmento não é

obrigatório, pois a fonte dos dados pode ser indicada externamente à sintaxe da consulta, no momento em que é realizada a execução do interpretador das consultas do SPARQL.

Por fim, o *where* apresenta a função de selecionar dentro de todo o *data set*, o conjunto de dados que satisfaça as operações construídas pelo usuário, e que será utilizado como base, para que a cláusula *select* selecione somente as variáveis desejadas. Dentro desse fragmento, é possível criar relações, inferências, agrupar os dados, realizar filtros, criar novas variáveis, entre outras diversas opções. Sintetizando, é dentro do *where* que será definido como a consulta será realizada.

A Figura 6 demonstra graficamente as funções realizadas pelas cláusulas *where* e *select*, sendo possível ver que o *where* selecionará, dentro do conjunto total de dados, um fragmento que atenda às suas expressões, e o *select* selecionará, dentre esse fragmento, as partes que lhe interessam.

Figura 6 - Funcionamento das cláusulas *where* e *select*.



Fonte: Ducharme (2011, p. 4).

Conforme dito com relação aos resultados obtidos por meio de uma consulta SPARQL, segue a estrutura determinada no *select*. No exemplo da Figura 5, a consulta deseja retornar o CPF, o nome e saldo bancário, dentro de um conjunto de dados chamado de “banco.ttl”. A cláusula *where* indica o relacionamento entre o CPF e o saldo bancário, indicando que o saldo apresentado deverá ser necessariamente o pertencente àquele CPF. Além disso, são inseridos dois prefixos, um para pessoa e outro para banco. O exemplo do resultado dessa consulta é apresentado no Quadro 5.

Quadro 5 - Exemplo de resultado da consulta apresentada na Figura 6

cpf	nome	saldo
666.726.447-25	José da Silva	245,65
028.984.385-53	Maria de Souza	2089,78
153.330.345-24	João Santos	-1290,02

Fonte: Elaborado pelo autor

No Quadro 5 é possível ver três colunas de dados, que são apresentadas como resultados, pois no *select* da consulta foram escolhidas as três variáveis, cpf, nome e saldo, para serem representadas.

O SPARQL tem uma relação com os processos de recuperação da informação semântica, por ser uma linguagem que explora a expressividade semântica do OWL e do RDF. No entanto, para explicar o uso do SPARQL na recuperação da informação, é necessário que seja feita uma explanação sobre as principais teorias desse campo de estudos. Na próxima seção serão abordadas tais teorias.

4 RECUPERAÇÃO DA INFORMAÇÃO

A recuperação de informação tem se tornado alvo de muitos estudos, devido a grande quantidade de informações disponíveis, nas mais variadas formas de organização em rede.

Em síntese, a recuperação de informação lida com a representação, armazenamento, organização e acesso às informações, devendo prover o usuário daquilo que ele necessita, de uma maneira facilitada (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nesse contexto, cabe fazer uma ressalva no que tange ao conceito de recuperação da informação frente à recuperação de dados. A recuperação de dados consiste em extrair de um banco de dados qualquer documento que contém uma expressão regular, ou os termos contidos ali, sendo que a recuperação de informação vai além, levando em conta a sintaxe e a semântica daquela informação, buscando satisfazer o que o usuário está pesquisando (BAEZA-YATES; RIBEIRO-NETO, 2013).

Dessa maneira, a recuperação de informação tem assumido um papel de destaque na Ciência da Informação, pois aparece como um elo na busca de encontrar a informação mais adequada ao usuário, no menor tempo possível.

O processo de recuperação de informação não consiste apenas em técnicas e métodos que envolvem o armazenamento e os algoritmos de recuperação, mas também em adaptar os sistemas no comportamento do usuário, entendendo, dessa maneira, como é a construção da informação e das instruções para a recuperação de informação. (SANTAREM SEGUNDO, 2010).

Com o surgimento da Web houve grande aumento no volume das informações eletrônicas, que trouxeram muitas vantagens quanto à possibilidade de troca, difusão e transferência de dados. Entretanto, este crescimento trouxe muitos problemas relacionados ao acesso, busca e recuperação das informações de real valor imerso em grandes volumes de dados (MODESTO, 2013).

Assim, um dos desafios da recuperação de informação é conseguir fazer com os ambientes informacionais digitais entendam o que o usuário está necessitando, de forma que os resultados vindos da busca possam ser de real valor e importância para o usuário, conseguindo dessa maneira ter uma maior aderência na intersecção entre os itens bibliográficos e as necessidades informacionais do usuário. (FUSCO, 2010).

Visando estudar e aprimorar tal intersecção, uma série de estudos estão sendo realizados discutindo um conceito de grande importância para a recuperação da

informação, que trata da relevância dos resultados extraídos de um conjunto de informações. A relevância apresenta diversas peculiaridades, em que são conduzidos estudos que partem da Ciência da Informação, tendo uma correlação indispensável com as áreas da Ciência da Computação.

Um pesquisador de renome dentro da Ciência da Informação, Hjørland, discute a questão da relevância dentro da Ciência da Informação, revelando diversas diferenciações no olhar que os pesquisadores devem possuir ao apontar a relevância. O autor, sintetizando suas discussões no texto supracitado, expõe 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)

O trecho citado expõe a dificuldade dos estudos tratando da relevância da informação, sendo necessário haver uma sintonia entre algoritmos e o sujeito que utiliza um programa. Ainda com essas variáveis, por vezes, criar sistemas que considerem a relevância se torna um trabalho custoso e complexo.

O conceito de relevância foi discutido anteriormente por outros autores, que embasaram o pensamento de Hjørland. Um destes autores é Borlund, que divide o conceito de 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 pensamento do autor demonstra como a relevância se aplica para a recuperação da informação, fazendo uma clara distinção entre a relevância segundo a ótica de um sistema e segundo a ótica de um usuário. O autor, no texto relatado, define uma série de critérios e de graus de relevância, sendo fundamentais para a aplicação de sistemas de recuperação da informação que visam aprofundar os estudos acerca da relevância.

A compreensão do pensamento destes dois autores (Hjørland e Borlund) demonstra a complexidade envolvendo a relevância e os processos de recuperação da informação. Há no Brasil algumas pesquisas que buscaram explorar esta temática, especialmente tratando de ambientes Web. Um trabalho que demonstra tal questão, aponta a dificuldade existente para se compreender a questão da relevância. Nesse sentido, 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 relevante ao menos relevante.

Ao expor essa questão, os autores trazem a necessidade de haver estudos que visem aprofundar a compreensão acerca dos processos de relevância, bem como do desenvolvimento de sistemas que tragam uma melhor noção de relevância para os usuários. No âmbito da recuperação da informação, os modelos clássicos podem auxiliar ao fornecer um bojo teórico amplo das técnicas e das tecnologias que podem auxiliar o tratamento dessa problemática. Diante do exposto, a seguir são apresentadas discussões acerca dos modelos de recuperação da informação.

4.1 Modelos de Recuperação de Informação

Os processos de recuperação de informação são baseados em modelos, que especificam os meios e as técnicas que são utilizadas. De acordo com Ferneda (2003, p. 18), “[...] a eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que o mesmo utiliza”

Os chamados modelos clássicos de recuperação de informação apresentam determinadas estratégias para as consultas. Nesses modelos, os documentos normalmente são representados como palavras-chaves, utilizadas como termos para a indexação (SANTAREM SEGUNDO, 2010). Os modelos chamados de clássicos são: modelo booleano, modelo vetorial e modelo probabilístico.

O modelo booleano é baseado na teoria dos conjuntos e possibilita a utilização dos operadores lógicos booleanos “E”, “OU” e “NÃO”, para a composição da expressão da busca, especificando os termos que devem ou não estar contidos nos documentos recuperados. O modelo tem esse nome, por ser baseado na álgebra de boole, criada por George Boole, que utiliza a linguagem binária, em que um dado pode ser somente verdadeiro ou falso.

Esse modelo apresenta uma desvantagem no que diz respeito à classificação dos documentos, pois o corpus, após a realização de uma busca, é dividido em dois subconjuntos: aqueles documentos que atendem à expressão de busca e aqueles que não atendem. Dessa forma, não existe um critério de classificação de relevância dos documentos, dificultando para que o usuário encontre os documentos que atendam mais

satisfatoriamente às suas necessidades informacionais (SILVA, SANTOS, FERNEDA, 2013)

O modelo vetorial, também chamado de espaço vetorial, funciona por meio do cálculo da similaridade entre um documento e a expressão de busca realizada pelo usuário. Por ser um modelo não binário, o cálculo irá mostrar qual é o nível de similaridade existente, o que é feito com o uso de vetores, e baseia-se na quantidade total de documentos, na quantidade de ocorrências de um termo no corpus de documentos e desse mesmo termo nos documentos individualmente. Com o uso desse modelo, ocorre uma evolução na questão da relevância, quando comparado ao modelo booleano, possibilitando, assim, realizar classificações e apresentar os documentos aos usuários em uma ordem, que possivelmente atenda mais adequadamente a suas necessidades informacionais (SOUZA, 2006).

Por fim o modelo probabilístico tem como princípio a probabilidade de um documento atender às necessidades do usuário. 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.

Esse modelo apresenta bons resultados após serem feitas as iterações, com um bom desempenho prático. Porém, não verifica a frequência dos termos dentro dos documentos, além de depender da precisão das estimativas de probabilidade (CARDOSO, 2004).

Outros modelos baseados nos clássicos e em teorias computacionais começam a ser utilizados, e podem agregar bastante ao processo de recuperação de informação. A utilização de modelos computacionais baseados em processos biológicos como algoritmos genéticos e redes neurais artificiais pode apresentar bons resultados quando unida à recuperação de informação, em que o sistema vai se adaptando ou aprendendo, conforme a utilização dos usuários e nos resultados que são recuperados (SANTAREM SEGUNDO, 2010).

Os três modelos clássicos apresentados demonstram parte do como se deu a evolução dos mecanismos de recuperação da informação, existindo atualmente uma gama de soluções que usam conceitos e técnicas de cada um desses modelos.

No que tange à relevância, as técnicas podem contribuir tanto do ponto de vista do usuário, quanto dos sistemas, pois o usuário, ao ter uma interação com o sistema, no

sistema probabilístico, permite que os algoritmos sejam capazes de obter informações acerca de suas necessidades, ao mesmo tempo que o modelo booleano será capaz de dividir o corpus de documentos entre aqueles que atendem ou não às necessidades informacionais dos usuários.

Como relatado, há diversas implementações que utilizam características distintas dos modelos apresentados. Tais implementações permitem tratar os dados com um nível adequado de eficiência, tanto no que diz respeito à velocidade de recuperação dos documentos, quanto à existência de critérios de relevância que combinam os diversos conceitos apresentados. No contexto desta pesquisa, foi pesquisado e utilizado uma implementação chamada de Apache Solr, descrita com mais detalhes na próxima subseção.

4.2 Apache SOLR

Os processos de recuperação de informação são executados por diversas ferramentas que implementam os modelos de recuperação de informação. Tais ferramentas apresentam distintas características, em que são utilizadas técnicas aprimoradas, com o intuito de executar uma melhor indexação e busca dos registros armazenados, visando melhor atender às necessidades informacionais dos usuários.

Essas ferramentas são implementadas em diversos softwares, que necessitam realizar a indexação e a busca de registros e documentos, de distintas fontes de informação e de variados formatos. Uma ferramenta de destaque é o Apache Solr, sendo uma ferramenta construída na linguagem de programação Java, estando atualmente em sua versão 5.4.0.

O site oficial da plataforma Apache Solr define-a como: “[...] a plataforma de código aberto de busca corporativa popular e rápida, construída sobre o Apache Lucene” (APACHE, 2014, tradução nossa).

Vale destacar que o Apache Lucene é uma biblioteca³ de indexação e buscas, construída em Java, que tem como característica permitir buscas com alta performance e alta velocidade de resposta, sendo adequado para qualquer aplicação que necessita realizar buscas em textos completos (APACHE, 2011-12).

³ Compreende-se por biblioteca um conjunto de códigos que permitem determinadas funções. Dessa forma, para que se utilize essa biblioteca, é necessário que seja implementado um programa que faça referência a ela, e utilize as funções por ela fornecida.

Assim, o Apache Solr é uma implementação do Apache Lucene, sendo um servidor que permite realizar indexação e consultas a documentos e registros, utilizando as principais funcionalidades de busca do Lucene.

No *website* do Solr, são apresentadas como as principais características de software: a capacidade avançada de buscas em textos completos; a otimização em alto volume de tráfego; os padrões abertos de interface: XML, HTML e HTTP; as interfaces de administração completas; o fácil monitoramento; ser altamente escalável e tolerante a falhas; ser flexível e adaptável com fácil configuração; a indexação próximo a tempo real; e a arquitetura extensível com *plug-ins* (APACHE, 2014).

Por meio destas características, é possível identificar que o Apache Solr foi construído para ser rápido e eficiente ao trabalhar com grandes volumes de dados. Percebe-se isso, ao verificar a flexibilidade existente durante a implantação do sistema, em que é possível adequar os formatos de acordo com os dados que são recebidos.

Vale destacar ainda que a característica de ser altamente escalável, torna viável a inserção de grandes quantidades de dados, que poderão ser processados rapidamente. Outro destaque fica por conta da indexação computacional, que ocorre próximo a tempo real, oportunizando aos ambientes digitais que utilizam o Solr uma boa atualidade, não havendo perdas de informações, relacionadas ao processo de indexação.

Ademais, existem alguns passos necessários para que seja possível utilizar a plataforma Apache Solr em qualquer sistema operacional. Primeiramente, o Java deve estar instalado, pois a plataforma funciona sobre esse sistema. Em seguida, são realizados o download e a instalação do Solr; no site oficial (<http://lucene.apache.org/solr>) é possível encontrar o link do arquivo para o download, e no manual de iniciação é explicado como devem ser feitas a instalação e a configuração do servidor em que estará funcionando a plataforma (<http://lucene.apache.org/solr/quickstart.html>).

Em seguida, é criada uma base de dados, onde ficarão armazenados todos os documentos que serão indexados. Tal base deverá ser configurada conforme os tipos de documentos que nela serão inseridos. Existem diversos tipos de configurações padrões e, também, a possibilidade de ser criada uma configuração personalizada, conforme as necessidades de cada sistema específico. Estando com a base de dados criada e configurada, basta serem inseridos documentos para a indexação computacional, e serem realizadas consultas que recuperem as informações.

O Solr apresenta um amplo espectro de possibilidades para a realização de consultas, permitindo buscas booleanas, buscas com lógica fuzzy, buscas por

proximidades, entre outras. Tal característica apresenta-se como primordial, pois possibilitará que os resultados obtidos a partir das buscas apresentem uma maior precisão e, como consequência, possam atender com mais eficiência às necessidades informacionais dos usuários.

Destacam-se as buscas booleanas, em que as expressões de buscas podem conter os operadores AND, OR e NOT, para melhor representar os termos que deverão ser recuperados a partir de uma consulta. Além disso, podem ser inseridos parênteses “(” e “)”, na montagem da busca, como em uma expressão booleana tradicional, permitindo, assim, a construção de expressões mais complexas.

Destacam-se ainda, os critérios de relevância utilizados por essa plataforma. Em suma, a relevância irá definir a ordem em que os resultados serão apresentados para os usuários, havendo uma compreensão de que os primeiros resultados, teoricamente, seriam capazes de melhor atender às necessidades informacionais dos usuários.

A ordenação desses resultados pode seguir diversos critérios; no entanto, o próprio Solr oferece uma forma de relevância que considera a proximidade dos termos de busca e dos registros armazenados. Esse critério é chamado de “relevancy score”, ordenando os resultados obtidos, em que os primeiros resultados são mais relevantes, enquanto os últimos são menos relevantes. Vale destacar que o Apache Solr permite que as ordenações dos resultados sigam outras regras, como a data dos documentos ou a ordem alfabética.

O critério utilizado é do Lucene, em que a busca ocorre por proximidade, levando em consideração critérios sobre as características dos documentos indexados computacionalmente, bem como dos campos apresentados por tais documentos, entre outros. No caso de registros de metadados em XML, esses critérios são definidos com bastante precisão, uma vez que é conhecida a estrutura de documento que será inserido no sistema. (APACHE, 2012).

Deve-se destacar, ainda, que a relevância não está limitada à ordenação dos resultados, havendo uma gama de variáveis que poderão ser utilizadas para acrescentar os critérios de padrões de relevância. No contexto desta pesquisa, a ordenação dos resultados ocorre por meio dos algoritmos do Apache Lucene, porém são inseridas informações contextuais dos dados que visam contemplar o domínio em que uma busca está ocorrendo, além das necessidades dos usuários, ao promover uma interatividade durante a busca.

Tais apontamentos estão diretamente envolvidos com a relevância, pois contemplam uma forma de entregar melhores resultados aos usuários. Diante do exposto,

na próxima seção será explicitada a maneira como o modelo de recuperação da informação baseado nas tecnologias da Web Semântica foi concebido e construído.

5 MODELO DE RECUPERAÇÃO DA INFORMAÇÃO PARA REPOSITÓRIOS BASEADO EM ONTOLOGIAS

Os repositórios digitais são ferramentas capazes de reunir grandes quantidades de documentos, o que revela sua importância, considerando-se que é um desafio para a área de recuperação da informação encontrar meios capazes de satisfazer com eficiência as necessidades informacionais dos usuários nesses ambientes.

Nesse contexto, as organizações e as instituições de ensino necessitam reunir em ambientes digitais as suas produções intelectuais, de modo a prover aos seus alunos, profissionais e professores, os conhecimentos produzidos sob a sua tutela. No momento, os repositórios digitais têm desempenhado essa função com êxito, proporcionando um ambiente de divulgação e com facilidade de acesso aos objetos digitais.

No que tange aos repositórios digitais, há um meio muito difundido que busca reunir em um único ambiente os objetos digitais espalhados em diversos repositórios, em sistemas abertos de publicação de periódicos, entre outros. Esse meio são os sistemas de interoperabilidade em repositórios digitais, que utilizam como base o protocolo OAI-PMH para realizar a coleta de metadados, por meio do processo de *harvesting*.

No entanto, ambientes de interoperabilidade de repositórios digitais têm como desafio apresentar uma recuperação eficiente, devido à grande quantidade de documentos que tais ambientes englobam. Ademais, a busca sintática, por vezes, não consegue compreender quais são as necessidades dos usuários, recuperando uma quantidade de documentos muito extensa, que, por vezes, faz com que os indivíduos não encontrem algum documento que lhes interessa.

Nessa perspectiva, a inserção de elementos que agregam semântica aos processos realizados na recuperação de informação pode aparecer como uma alternativa para contextualizar o domínio em que o usuário realiza a busca. Um modo de realizar tal contextualização se dá por meio da utilização de tecnologias da Web Semântica, como as ontologias construídas em OWL, trabalhando com o auxílio da linguagem de consultas SPARQL, e a representação de dados com o RDF, que conseguem montar toda uma plataforma semântica de recuperação de informação.

Uma outra questão que pode ser considerada em ambientes de interoperabilidade em repositórios digitais diz respeito à possibilidade do usuário escolher as fontes informacionais utilizadas durante o processo de busca. A necessidade desse processo interativo, em que o usuário escolhe os repositórios utilizados na busca, pode ser

identificada devido ao fato de os provedores de serviços delimitarem os repositórios de onde foram extraídos os metadados anteriormente, não permitindo que novas fontes sejam inseridas diretamente pelo usuário.

Dessa forma, o presente modelo possibilita que novos repositórios possam ser cadastrados, para que seja realizado o processo de *harvesting*, permitindo que os metadados de um novo repositório sejam utilizados durante a busca. Assim, o usuário poderá escolher os repositórios que sejam de seu interesse, além de retirar da busca, aqueles que não lhe interessam. Essa possibilidade insere um nível alto de personalização de busca, pois dá ao usuário o controle das fontes utilizadas para encontrar os objetos digitais.

Dentro do contexto de recuperação de informação em repositórios digitais, foi proposto um modelo que tem como objetivo promover, mediante um sistema interoperável de repositórios digitais, uma recuperação de informação auxiliada pelo uso de conceitos e tecnologias da Web Semântica.

Tal modelo é central nesta dissertação, pois é por meio dessa proposta que se visa atender aos objetivos descritos e a responder a problemática exposta. Em suma, o modelo tem como função contextualizar a busca, por meio do uso de ontologias, que são capazes de expressar de modo computacional um determinado domínio.

Essa contextualização ocorre com uma relação desenvolvida neste trabalho, em que se identifica como cada propriedade do OWL pode ser utilizada no âmbito da recuperação da informação, permitindo que uma ontologia seja explorada para localizar relações que expressem o domínio da busca, além de identificar a semântica formal que uma relação possui com um determinado termo de busca, permitindo a criação de novas expressões de busca, com uma maior expressividade.

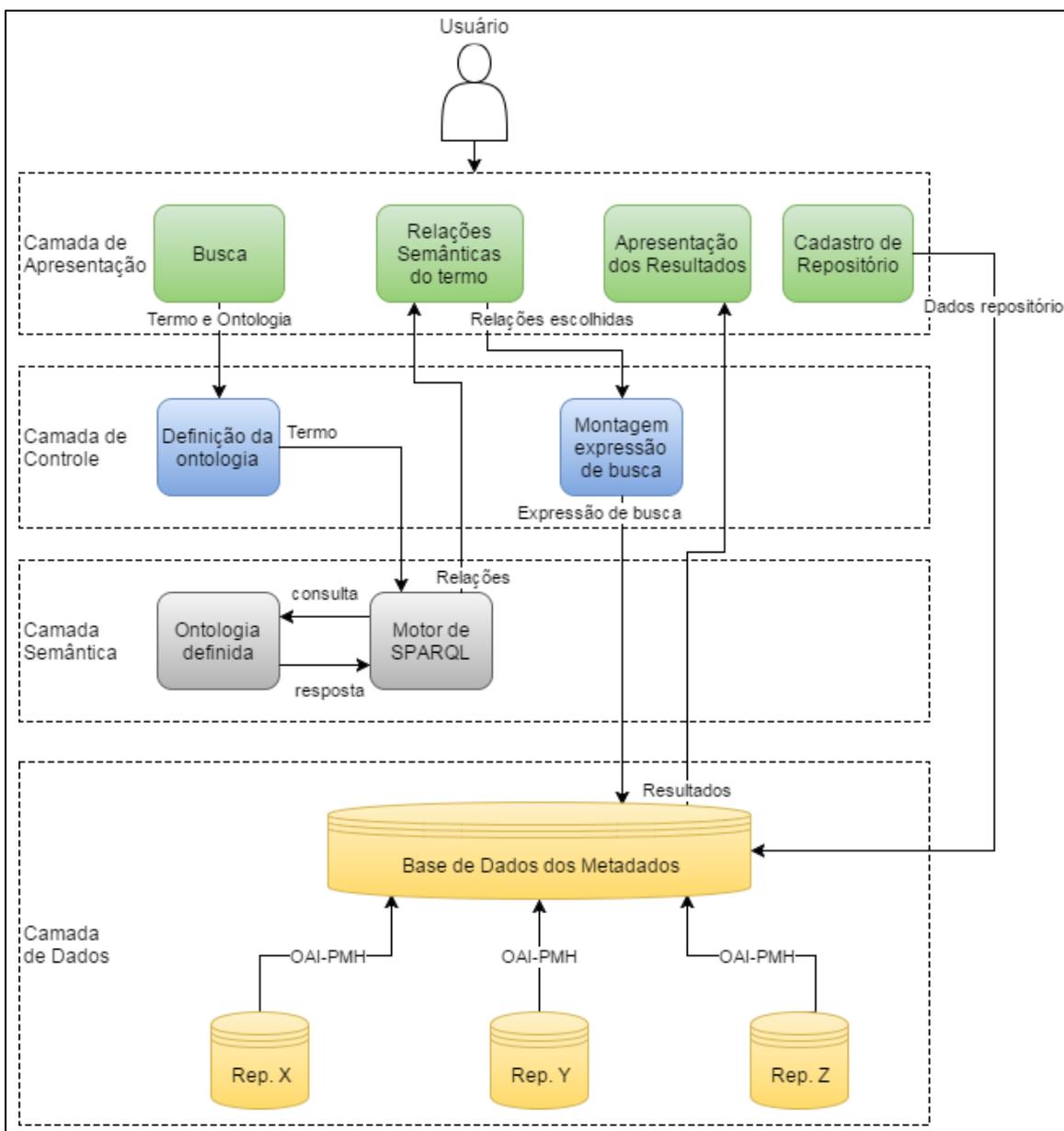
O processo de criação de novas expressões de busca contempla ainda uma interação com o usuário, em que este deve selecionar, dentre as relações encontradas na ontologia, aquelas que estão de acordo com suas necessidades informacionais. Essa característica dá ao usuário o controle dos termos utilizados, evitando que sejam utilizados termos que não estejam no contexto do usuário, dando mais interatividade ao modelo.

Outro ponto abordado pelo modelo foi relatado anteriormente; é o caso em que o usuário tem a possibilidade de escolher as fontes informacionais a serem utilizadas, bem como inserir outros repositórios que não estão cadastrados no sistema. Por meio dessa opção, o modelo permite uma expansão no nível de interatividade do sistema,

complementando a interação relatada anteriormente, em que o usuário escolhe as relações encontradas na ontologia de acordo com suas necessidades.

Diante do exposto, o modelo construído abordando as características supracitadas, pode ser visualizado na Figura 7.

Figura 7 - Modelo Semântico de Interoperabilidade entre Repositórios



Fonte: Elaborado pelo autor

No modelo proposto, o usuário realizará uma busca em uma base de dados que reúne a produção de diversos repositórios digitais. Tal usuário terá acesso a uma interface de buscas, que permitirá a ele a execução de duas tarefas: 1) a realização de uma busca e

2) o cadastro de novas fontes informacionais. Em síntese, os usuários desse modelo poderiam ser estudantes e pesquisadores, visando localizar, dentro de repositórios digitais, produções científicas e intelectuais.

A realização de uma busca consiste, primeiramente, na escolha, por parte do usuário, de uma ontologia que represente o domínio em que sua busca se encontra. Na sequência, ocorrerá a definição de uma expressão de busca, em que o usuário irá digitar na interface, termos que representem a sua necessidade informacional, bem como escolherá em quais dos repositórios cadastrados será realizada a busca. Posteriormente, o sistema identificará as relações semânticas que o termo digitado pelo usuário possui, utilizando como base a ontologia definida anteriormente.

Com as relações identificadas, o usuário escolherá, dentre as relações encontradas, aquelas que estão de acordo com a sua busca, sendo realizada, na sequência, a montagem de uma nova expressão que engloba as características semânticas das relações escolhidas. Utilizando a expressão construída, realiza-se a busca na base de dados de metadados, daqueles documentos que atendam à expressão, retornando os resultados aos usuários.

O segundo processo é o do controle das fontes informacionais. Em síntese, esse processo consiste na possibilidade de serem inseridas novas fontes informacionais no processo de recuperação de informação. Esse procedimento deve ser utilizado no caso de um repositório que o usuário deseje utilizar em sua busca não estar listado no sistema, sendo permitido, então, a inserção desse repositório, realizando o *harvesting* dos metadados.

Outra característica apresentada na Figura 7 é a existência de quatro camadas principais no modelo: camada de apresentação, responsável pela interação visual entre o usuário e o sistema, permitindo que o usuário possa inserir a expressão de busca e escolher as ontologias e os repositórios que serão as fontes informacionais; camada de controle, responsável por gerenciar os processos de buscas, além de ser uma interface entre o usuário e a ontologia; camada semântica, que trata dos processos relacionados à contextualização do termo de busca escrito pelo usuário, dentro de uma ontologia e camada de dados, que representa as fontes informacionais, repositórios, que trabalharão como provedores de dados, para a realização das buscas.

As quatro camadas apresentadas serão detalhadas nas subseções seguintes.

5.1 Camada de Apresentação

A camada de apresentação possui um papel central no modelo, pois fica a cargo dela a interação que o usuário terá com o sistema. Dessa forma, torna-se necessário que esteja claro quais serão os momentos em que o usuário interagirá com o sistema, bem como as ações que deverão ser realizadas.

A partir do modelo construído, verificamos a existência de quatro momentos, que serão descritos com maiores detalhes na sequência: 1º) a escolha da ontologia, dos repositórios que serão utilizados como fontes e a escrita dos termos que representem a necessidade informacional do usuário; 2º) a escolha das relações encontradas na ontologia, dentre as relacionadas com o termo escolhido pelo usuário; 3º) a apresentação dos resultados que atenderam à expressão de busca construída pelo sistema; e 4º) o cadastramento de uma nova fonte informacional para a camada de dados.

O primeiro momento possibilita que o usuário insira uma ontologia que contextualize o domínio da sua busca, sendo que tal ontologia deverá ser estruturada na linguagem OWL ou RDF *Schema*. A inserção de uma ontologia consiste em que o usuário encontre uma ontologia pronta e a deposite no sistema, de modo que o sistema irá identificar e processar tal artefato. Além disso, o usuário deverá ter acesso a uma lista com os repositórios que estão contidos na camada de dados, para que possa escolher quais desses serão utilizados como fontes informacionais da busca. Por fim, o usuário deverá digitar os termos que representem a sua necessidade informacional, para que possa ocorrer a recuperação da informação.

O segundo momento diz respeito ao processo de escolha das relações semânticas do termo de busca, que foram encontradas na ontologia. A interface receberá da camada de controle uma lista com as relações semânticas, contendo os termos relacionados e as suas respectivas propriedades semânticas. Essa lista será apresentada ao usuário, que escolherá aquelas que estão de acordo com a sua busca.

O momento posterior representa os resultados que atenderam à expressão de busca construída pelo sistema; nesse momento o usuário poderá ter acesso aos documentos e aos seus metadados, além de ter a possibilidade de acessar o link externo, que conduz ao objeto digital original, arquivado no repositório de origem.

O último momento é a possibilidade do usuário cadastrar um novo repositório como fonte informacional. Nesse processo, o usuário deverá inserir os dados do nome e da URL de um repositório que utilize o OAI-PMH, para que o sistema possa realizar o processo de *harvesting* dos metadados. Como consequência dessa inserção, o usuário, ao

realizar uma nova busca, poderá escolher esse novo repositório cadastrado, como uma fonte informacional para a sua busca.

Os processos explicitados apresentam uma visão sintetizada do sistema, uma vez que contemplam somente a interação entre usuários e sistema. Todas as fases citadas são compostas por uma série de procedimentos, descritos nas próximas subseções.

5.2 Camada de Controle

A camada de controle funciona como uma relação entre a interface visual com a camada semântica e com a camada de dados. Essa camada terá como função a execução de três tarefas, que serão descritas na sequência.

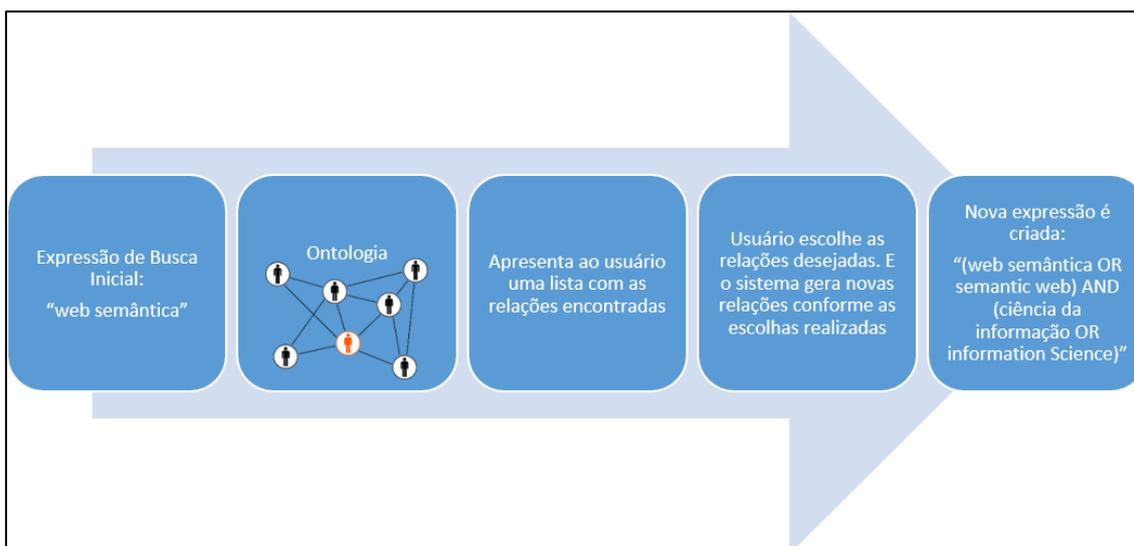
Primeiramente, a camada de controle deverá solicitar à camada semântica, as relações semânticas encontradas. Tal processo ocorre posteriormente à definição da ontologia escolhida, do termo de busca e das fontes informacionais pelo usuário, descritos anteriormente na camada de apresentação. A resposta fornecida, da camada semântica, com as relações que o termo de busca possui, será enviada à camada de apresentação, em que o usuário realizará a escolha das relações que estão convergentes com as suas necessidades informacionais. Posteriormente, o controle montará a expressão de busca, embasado nas relações escolhidas.

O processo da montagem da nova expressão de busca não é uma responsabilidade do usuário, pois essa montagem ocorrerá automaticamente, de acordo com as relações escolhidas por tal usuário. A montagem ocorre conforme o significado semântico que cada propriedade escolhida possui. A especificação de como cada propriedade atuará dentro desta montagem, será dada na subseção da camada semântica, em que será especificado como o OWL atua nos processos de recuperação da informação.

Tal processo ocorre pelas especificações que as relações de ontologias construídas em OWL possuem, como por exemplo subclasses, classes equivalentes, propriedades transitivas, classes inversas, classes similares, entre diversas outras propriedades que podem ser utilizadas para a construção de novas expressões de busca.

Um exemplo teórico de como este processo de montagem de expressão de busca ocorre pode ser visto na Figura 8, em que estão sintetizados os passos para a construção de uma nova expressão de busca, considerando as relações da OWL escolhidas pelos usuários.

Figura 8 - Passos para a construção da expressão de busca automática



Fonte: Elaborado pelo autor

Nesse exemplo demonstrado pela Figura 8, o usuário busca por Web Semântica, sendo retornado para ele um conjunto de relações. Nessas relações, foi entendido como sendo uma relação de igualdade de Web Semântica com *Semantic Web*, e relacionamentos de proximidade com Ciência da Informação e *Information Science*. O modelo poderá mostrar para o usuário relações de igualdade com ontologia e com Web 3.0, por exemplo; contudo, tal usuário pode não querer relacionar tais termos, de forma que não os escolhe para aprimorar sua busca. Assim, é possível verificar que a lista de relações será utilizada como uma ferramenta que agregue especificidades, reduza ambiguidades e melhore a qualidade da expressão de busca.

Por fim, estando com a expressão de busca, é realizada a solicitação dos documentos que atendam a tal expressão. A solicitação é realizada para a camada de dados, que retorna os documentos que atenderam à expressão construída, entregando à camada de visualização uma lista com tais documentos.

A relação entre a camada de controle e a camada semântica concerne às questões relativas à escolha dos termos pelos usuários e à montagem da expressão de busca, considerando as características semânticas da OWL. Desta feita, a próxima subseção explora as questões relativas à camada semântica.

5.3 Camada Semântica

A camada semântica é responsável pela contextualização da busca feita pelos usuários, visando compreender a característica do domínio em que os termos de busca e

os objetos digitais se encontram. Para isso, o centro dessa camada é uma ontologia escolhida pelo usuário, que, em conjunto com outros mecanismos baseados nas tecnologias e nos conceitos da Web Semântica, é capaz de verificar as relações que um determinado termo possui.

Nessa perspectiva, as ontologias permitem que o contexto e o significado dos termos possam ser encontrados, para que assim, um mecanismo de recuperação da informação possa aprimorar a localização das informações que melhor atendam às necessidades informacionais dos usuários. Um meio de explorar tais ontologias é a utilização da linguagem de consultas SPARQL, capaz de percorrer conjuntos de dados expressos em OWL e RDF.

Partindo dessa ideia, a camada semântica está centrada em dois elementos: a ontologia e o motor de SPARQL. A seguir serão explicadas a função e as características desses dois elementos, no que concerne a este trabalho.

A ontologia tem a função de contextualizar a busca, sendo o elemento que permitirá, aos mecanismos computacionais, verificar as relações e as propriedades que um termo possui. No âmbito desta pesquisa, o usuário tem a função de definir a ontologia que será utilizada, no momento de inserir os termos de busca, sendo que este modelo apresenta suporte às ontologias construídas em OWL e RDF *Schema*.

Em contrapartida, o motor de SPARQL deve identificar as relações que um determinado termo possui dentro da ontologia escolhida pelo usuário. Este instrumento tem como princípio a capacidade de identificar todos os tipos de relacionamentos e de propriedades que um termo possui.

Há ainda um segundo princípio que centra este motor, que corresponde à capacidade de utilizar os elementos semânticos da OWL referentes às propriedades. Essa possibilidade faz com que o motor de SPARQL descubra uma quantidade grande de informações relacionadas a um determinado termo, sendo que tais informações seriam extraídas do contexto de onde elas foram obtidas.

Em suma, o motor de SPARQL, ao explorar as relações existentes na ontologia para localizar os termos relacionados à busca do usuário, é capaz de identificar propriedades semânticas dessas relações, permitindo que sejam obtidas informações mais precisas das classes relacionadas.

Os dois processos do motor de SPARQL, o primeiro referente à localização das classes relacionadas e o segundo referente às propriedades semânticas das relações, utilizarão como princípio a geração de consultas SPARQL em tempo de execução. Essa

característica significa que a partir do instante em que o usuário define um termo e uma ontologia, o motor de SPARQL deverá criar as consultas que possibilitem a identificação das relações existentes.

O fato da geração de consultas ocorrer em tempo de execução mostra-se como o ponto central e de maior complexidade desse processo, pois como não há um conhecimento prévio da ontologia, é necessário que as consultas se adaptem a cada estrutura escolhida pelo usuário.

Tais fatores têm como objetivo central localizar informações significativas que contextualizem o termo de busca. Dessa forma, há a necessidade de serem buscadas relações que ultrapassem a barreira das relações hierárquicas, utilizando as propriedades da OWL, como em propriedades de axiomas e outras propriedades especiais.

Na prática, o motor de SPARQL realiza três processos sequenciais: o mapeamento da ontologia; a localização de um termo dentro da ontologia; e a identificação das relações do termo de busca na ontologia. Na sequência serão descritos esses três processos, com detalhes.

Primeiramente, é realizado o mapeamento da ontologia, em que o usuário escolhe uma ontologia, que pode estar disponível na Web, ou localmente na máquina do usuário, e o sistema deverá realizar o mapeamento computacional, tornando-a uma estrutura em que possam ser realizadas consultas SPARQL. Tal mapeamento consiste de técnicas de programação para receber um arquivo em OWL ou em RDF e convertê-lo em classes da linguagem de programação.

O segundo passo trata da localização de um termo desejado dentro da ontologia mapeada anteriormente. Para realizar essa localização, é gerada uma consulta SPARQL, que deverá encontrar a posição do termo na ontologia. Localizar o termo dentro da ontologia é essencial para a continuidade do processo, pois é necessário que seja conhecido do sistema a URI que um termo possui, para que, assim, possam ser criadas as próximas consultas SPARQL. Para tal, o gerador necessita consultar o mapeamento da ontologia que fora realizado. Caso o termo seja localizado dentro da ontologia, será identificado a URI, com o seu prefixo e a sua sintaxe.

O último passo trata da identificação das relações existentes. Tal procedimento se dá por meio de consultas SPARQL, que são geradas analisando as ligações que o termo de busca possui. Esse processo ocorre após a localização do termo, iniciando uma série de consultas sucessivas à ontologia, buscando encontrar as relações existentes entre o termo do usuário e os outros elementos da ontologia. As consultas SPARQL apresentam

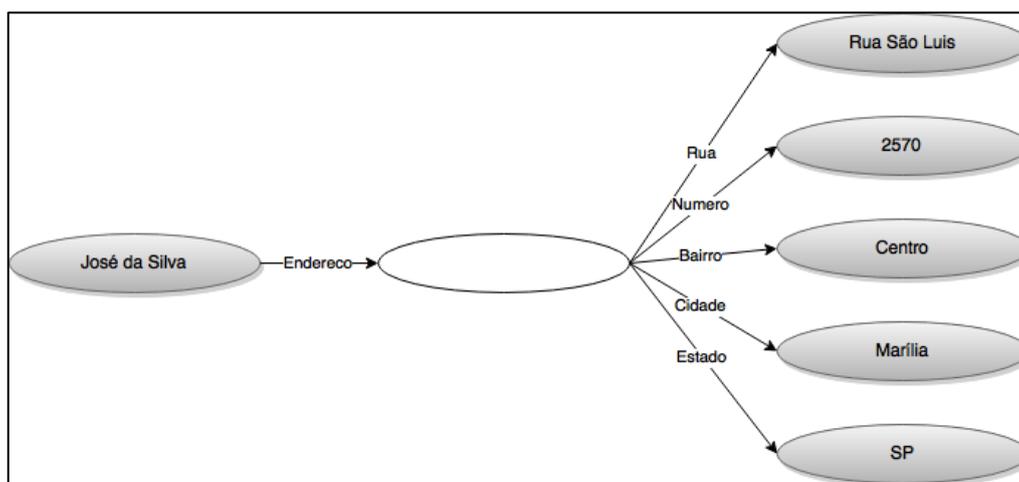
a seguinte estrutura de triplas RDF: o sujeito do RDF indica o elemento que corresponde ao termo principal, sendo que o predicado e o objeto são elementos variáveis, que permitirão encontrar as relações existentes entre o sujeito (termo principal) com o restante da ontologia.

Nessa fase, a identificação de propriedades especiais da OWL terá a função de auxiliar na localização de ligações entre algumas classes e as suas respectivas propriedades. Quando o motor de SPARQL se deparar com algumas propriedades, será possível inferir o significado que tais relações possuem, permitindo a descoberta de outras classes relacionadas.

Não obstante, há outras questões a serem tratadas pelo motor de SPARQL, que aumentam significativamente a complexidade dos processos realizados. Dentre elas, a existência dos chamados nós em branco dentro do RDF, exige o tratamento diferenciado dos processos.

Os nós em branco são estruturas em RDF, em que uma classe tem relacionamento com diversas outras classes. 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, como pode ser visto na Figura 9.

Figura 9 - Exemplo de grafo com nó em branco



Fonte: Elaborado pelo autor

Este mecanismo do RDF apresenta um alto nível de complexidade no processo de geração automática de consultas, pois não apenas é desconhecida a quantidade deles, como também não se sabe quais são os nós em branco que existem em um relacionamento. Além disso, um nó desse tipo pode apontar para diversos nós, que, por sua vez, podem apresentar algum nó em branco e, assim, sucessivamente, sendo necessário que o gerador crie consultas SPARQL que vão percorrendo todos os nós, para que sejam extraídos os relacionamentos que um determinado objeto possui. Os nós em branco não podem ser desprezados, pois podem existir informações de grande importância relacionadas a eles.

Dessa forma, tendo o gerador de consultas SPARQL tratado de todas essas questões, ele verificará todos os relacionamentos que o termo buscado pelo usuário possui. Ao final, haverá uma lista com todos os relacionamentos existentes entre o termo principal e outros objetos da ontologia.

Por fim, a lista com os relacionamentos é enviada ao modelo principal, para que seja apresentado ao usuário as relações encontradas e, assim, ele possa selecionar aquelas que estão relacionadas a suas necessidades informacionais.

Os procedimentos apresentados na camada semântica possuem relação direta com a ontologia e com o termo de busca escolhido pelo usuário, o que permite contextualizar tal termo dentro de um domínio. A capacidade semântica do modelo relaciona-se diretamente com as propriedades existentes na OWL, havendo a necessidade de descrever como os tipos de propriedades de um relacionamento influenciam na criação de uma expressão de busca. Na próxima seção será relatado como ocorre esse relacionamento, identificando a relação entre as propriedades da OWL e a recuperação da informação.

5.3.1 OWL na Recuperação de Informação

A linguagem de construção de ontologias OWL possui uma série de elementos que permitem a inserção de informações para a contextualização de um determinado domínio. No subcapítulo deste trabalho denominado de “OWL”, são apresentados diversos elementos de uma ontologia, que são capazes de contextualizar um determinado termo, por meio de propriedades que agregam na descrição, como visualizado no quadro 4.

No contexto do modelo proposto neste trabalho e, mais amplamente, de recuperação da informação, é necessário explicitar como tais propriedades podem ser utilizadas na prática, buscando manusear as diversas características que uma ontologia

possui, não estando restritas a características inerentes a taxonomias ou a tesauros. Relacionar ontologias, mais especificamente a linguagem OWL, com a recuperação da informação, possibilita apontar como as propriedades semânticas desses artefatos podem ser utilizadas no processo de construção de uma expressão de busca, sendo tal questão de suma importância para haver um avanço na concepção do uso real de ontologias para recuperação da informação.

Visando definir as relações existentes entre as ontologias OWL e a recuperação da informação, utilizou-se dos quadros construídos nas subseções de RDF e OWL, quadros 3 e 4, como base na descrição de como as propriedades deverão ser utilizadas no processo de recuperação da informação. Utilizaram-se as propriedades do RDF *Schema* e OWL, pois ambas trabalham em complemento, sendo que as ontologias em OWL necessariamente utilizam as classes e propriedades do RDF *Schema*.

Por meio das propriedades apresentadas pelas tabelas, verificou-se a existência de duas classificações principais das relações existentes: 1) relações que tratam das propriedades da ontologia e 2) as relações que tratam das classes da ontologia. Dentro de cada uma dessas classificações, há uma série de divisões que apontam o comportamento das relações e das propriedades dentro dos processos de recuperação da informação.

Na sequência são apresentados dois quadros que classificam as propriedades do OWL dentro das duas categorias supracitadas.

Tais quadros apresentam três colunas: a primeira contendo os nomes das propriedades, a segunda contendo uma divisão entre as propriedades daquele grupo, em que se verifica um mesmo comportamento para a recuperação da informação, e a terceira coluna com as características dessas propriedades, no contexto da recuperação da informação.

Primeiramente, o Quadro 6 traz as propriedades da OWL que descrevem as classes da ontologia. Nesse quadro não foram consideradas as propriedades “rdf:type”, “rdfs:label” e “rdfs:comment”, pois tais propriedades não possuem uma relação de contextualização relevante para a recuperação da informação, ou seja, não contribuem para apresentar o contexto que determinadas informações possuem.

Quadro 6 - Comportamento para recuperação da informação para propriedades de classes da OWL

Propriedade	Divisão	Características
<i>rdfs:range</i>	Aproximação	Os relacionamentos que são classificados como aproximação sintetizam as relações entre duas classes da OWL. No âmbito da recuperação da informação, as duas propriedades podem contextualizar os termos de busca, uma vez que identificam os relacionamentos de um determinado termo. Desta feita, as relações deste tipo correspondem à inserção de novos elementos na busca, que, caso o usuário ache necessário, deverá ser utilizada como uma intersecção entre os conjuntos dos resultados que possuem o termo de busca e os que possuem o termo relacionado.
<i>rdfs:domain</i>		
<i>rdfs:subClassOf</i>	Hierarquia (ou Especificação)	As propriedades de hierarquia apontam uma especificação ou uma generalização de uma determinada classe, ou seja, esse tipo de relação define um conceito mais genérico ou mais específico que uma classe possui. Tratando de recuperação da informação, esses termos podem auxiliar a localizar documentos, caso os resultados obtidos não sejam satisfatórios, ou sejam em uma baixa quantidade. Assim, é necessário que os processos de recuperação da informação façam uma união entre os resultados que possuem o termo de busca e os que possuem os termos relacionados por essas propriedades.
<i>owl:oneOf</i>	Relacionamento entre diversos conceitos relacionados	Esta propriedade promove a ligação entre diversas classes, pois indica que uma classe faz parte de um conjunto de classes relacionadas. Dessa forma, a propriedade de enumeração (<i>oneOf</i>) traça um relacionamento entre uma classe com diversas outras

Propriedade	Divisão	Características
		relacionadas. No que tange à recuperação da informação, essas propriedades podem apontar classes relacionadas, podendo auxiliar no contexto da busca. Para utilizar essas propriedades, seria possível traçar a intersecção entre os resultados obtidos do termo de busca e o conjunto obtido com a união dos resultados dos termos relacionados.
<i>owl:equivalentClass</i>	Igualdade	As classes de igualdade apontam termos e relações que apresentam o mesmo significado semântico. Dessa forma, no contexto da recuperação da informação, as relações existentes entre os termos de busca apontam outros termos que, caso estejam presentes nos resultados da busca, podem fornecer respostas satisfatórias. Para isso, essas propriedades podem ser utilizadas para fazer uma união entre o termo inicial e os termos relacionados.
<i>owl:sameAs</i>		
<i>owl:intersectionOf</i>	Intersecção	A intersecção aponta objetos oriundos de uma mesma classe, tendo, assim, algumas bases semânticas semelhantes. Desta feita, tratando de recuperação da informação, dois conceitos que são uma intersecção podem possuir uma ligação em que o processo de busca deverá fazer uma intersecção entre os resultados em que aparece o termo de busca e os resultados com o termo relacionado.
<i>owl:unionOf</i>	União	A união apresentará todos os objetos que são instâncias de pelo menos uma classe. Dessa forma, para a recuperação da informação pode haver laços comuns entre dois objetos relacionados por essa relação, contudo, essa relação não necessariamente será entre classes com a mesma origem, podendo ser instâncias de mais de uma classe. Assim, o processo

Propriedade	Divisão	Características
		de busca deverá realizar uma união entre os resultados que contêm o termo de busca e os que contêm o termo relacionado.
<i>owl:disjointWith</i>	Diferenciação	As seis propriedades apresentadas são enquadradas com uma mesma característica: de diferenciar os objetos relacionados, de modo que possibilitem inferir que duas classes apresentam essencialmente significados distintos. Na esfera da recuperação da informação, essas propriedades apontam que uma determinada relação entre termos possui significados semânticos diferentes, e que, portanto, poderia ser utilizada para excluir os resultados de busca que apresentem os termos de diferenciação, pois tratam de questões distintas.
<i>owl:AllDisjointClasses</i>		
<i>owl:disjointUnionOf</i>		
<i>owl:differentFrom</i>		
<i>owl:AllDifferent</i>		
<i>owl:complementOf</i>		

Fonte: Elaborado pelo autor.

As relações apresentadas pelo Quadro 6 apontam a diversidade de características que as propriedades da OWL possuem, havendo sete divisões dentre as propriedades que auxiliam na descrição das classes das ontologias. Vale destacar ainda que, na coluna das características das divisões, foi possível identificar como as propriedades estão relacionadas e podem ser utilizadas no âmbito da recuperação da informação.

As propriedades de classes, no contexto da recuperação da informação, podem ser utilizadas para apontar como um termo de busca se relaciona com outros termos. Em suma, ao identificar os termos relacionados a um determinado termo, é possível construir expressões de busca que sinalizam para um sistema de recuperação da informação os termos relacionados, com um nível de semântica que extrapola os relacionamentos de tesouros ou de taxonomias.

Nesse sentido, na busca de transformar os comentários tecidos no quadro 6 em uma linguagem computacional, capaz de ser traduzida em algoritmos, elaborou-se um segundo quadro, apresentando como um termo de busca se relaciona a um termo

identificado na ontologia, de acordo com a propriedade que une esses dois termos, utilizando operadores booleanos.

Para definir qual será o comportamento para cada propriedade foram utilizadas as divisões realizadas, em que são aglutinadas propriedades com ações semelhantes dentro da recuperação da informação, apresentadas no Quadro 7. Além disso, na terceira coluna é apresentada a ordem em que cada ação deve ser tomada, na construção da expressão de busca.

Quadro 7 - Ações das propriedades de classes

Divisão	Ação	Ordem
Aproximação	$((X) \text{ AND } Y)$	2
Hierarquia (ou Especificação)	$((X) \text{ OR } Y)$	1
Relacionamento entre diversos conceitos relacionados	$((X) \text{ AND } (Y1 \text{ OR } Y2 \text{ OR } Y3... \text{ OR } Yn))$	3
Igualdade	$((X) \text{ OR } Y)$	1
Intersecção	$(X) \text{ AND } Y$	2
União	$(X) \text{ OR } Y$	1
Diferenciação	$((X) \text{ NOT AND } (Y))$	4

Fonte: Elaborado pelo autor.

As ações apresentadas no quadro 7 correspondem a como o termo de busca (X) se corresponde aos termos encontrados na ontologia (Y). Vale destacar que no caso dos relacionamentos entre diversos conceitos relacionados, podem ser localizadas diversas classes; por esse motivo que a fórmula dessa divisão contém inúmeros termos (Y1, Y2, Y3 ... Yn).

Por meio das ações apresentadas, é possível criar uma expressão de busca compreensível computacionalmente e que leva em consideração o domínio em que um determinado termo está inserido. Nesse contexto, é possível que existam diversas classes relacionadas, o que conduz à necessidade de traçar a ordem em que uma expressão de

busca deve ser construída, caso existam diversas propriedades de classes. Dessa forma, a terceira coluna do quadro 7 indica a sequência de ações.

A ordem dessas ações foi assim elaborada, buscando manter o efeito que cada operação booleana possui no âmbito da expressão de busca. Por esse motivo, as operações centradas na expressão “OR” foram inseridas como as primeiras, devido ao fato de que recuperará os documentos que tem um termo ou outro. Na sequência serão inseridos os termos relacionados com o operador “AND”, que terá como efeito recuperar somente os documentos que possuem ambos os termos. Posteriormente, são relacionados os diversos termos relacionados, agregando os operadores “AND” e “OR”, com o mesmo efeito apresentado. Por fim, os operadores de negação, pois serão excluídos os resultados que apresentarem alguns termos recuperados, porém indesejados.

Vale destacar que o princípio interativo do trabalho propõe que o usuário poderá escolher quais os termos que serão utilizados na montagem de busca; no entanto, os tipos de relações serão obtidos automaticamente a partir das ontologias. Essa característica permite que os motores de busca não considerem termos que, embora apresentados pela ontologia como relacionados, não se adequam ao contexto do usuário.

Como relatado anteriormente, há um segundo tipo de propriedade da OWL que apresenta grande importância para o processo de recuperação da informação, são as propriedades de propriedades. Estas propriedades têm como função inserir características semânticas nas propriedades que relacionam duas classes de uma ontologia. Em síntese, elas buscam contextualizar uma relação, inserindo elementos que aprimorem a recuperação dessas relações futuramente.

O uso dessas propriedades pode ser amplamente explorado com o uso da linguagem SPARQL, em que é possível utilizar as propriedades da OWL para encontrar as relações e as classes existentes com mais eficácia. Tratando-se do cenário da recuperação da informação, essas propriedades podem aprimorar significativamente a localização dos termos, corroborando com o processo descrito anteriormente, que trata das propriedades de classes da OWL.

Dessa forma, a utilização de uma ontologia e do motor de SPARQL está condicionalmente ligada ao uso das propriedades tanto das classes quanto das propriedades. Demonstrando quais são as características das propriedades de propriedades na recuperação da informação, o quadro 8 indica as propriedades em si, uma divisão em que elas apresentam comportamento semelhante e as características supracitadas. Essa definição busca ser um apoio para o motor de SPARQL, para

possibilitar a descoberta das relações existentes de um determinado termo de busca, identificando classes relacionados, com o uso das propriedades da ontologia.

Cabe ressaltar que algumas propriedades foram excluídas para a montagem desse quadro, sendo elas: “*owl:propertyChainAxiom*”, “*owl:FunctionalProperty*”, “*owl:Inverse FunctionalProperty*”, “*owl:ReflexiveProperty*”, “*owl:IrreflexiveProperty*” e “*owl:NegativePropertyAssertion*”, pois elas, por mais que agreguem semântica na descrição de um domínio, não possuem uma ação concreta no âmbito da recuperação da informação.

Quadro 8 - Comportamento para recuperação da informação para propriedades de propriedade do OWL

Propriedade	Divisão	Características
<i>rdfs:subProperty Of</i>	Hierarquia	A hierarquia entre as propriedades aponta que uma propriedade filha herda as características da propriedade pai. Dessa forma, durante a descoberta das classes relacionadas em um processo de recuperação da informação, será possível identificar características semânticas das propriedades por meio da verificação de hierarquia entre as classes. Assim, ao identificar uma propriedade pai, por exemplo, seria possível identificar algumas características da OWL que uma propriedade filha possui, possibilitando encontrar outras classes relacionadas.
<i>owl:Transitive Property</i>	Relacionamento entre diversos conceitos relacionados	Esta propriedade permite localizar diversas classes relacionadas com um mesmo tipo de relacionamento, princípio este da transitividade. Para a recuperação da informação, esta propriedade possibilita que sejam encontradas classes que não estão diretamente ligadas a um termo principal, porém que possuem uma relação com outros termos que estão ligados a esse primeiro por uma mesma propriedade. Dessa forma, é possível traçar um conjunto de classes relacionadas a um

Propriedade	Divisão	Características
		determinado termo, que não possui uma relação direta com ele, mas que está ligado a outros termos por meio dessa mesma propriedade transitiva.
<i>owl:equivalente Property</i>	Igualdade	A existência de duas propriedades equivalentes permite que a descoberta de classes seja mais eficiente, uma vez que será possível identificar propriedades com o mesmo significado, o que pode evidenciar uma relação semelhante. Dessa forma, para a recuperação da informação, o uso dessa propriedade possibilita localizar relações entre classes semelhantes, podendo auxiliar na descoberta de classes relacionadas, com mais precisão.
<i>owl:Symmetric Property</i>	Simetria	A propriedade de simetria apresenta uma característica distinta das propriedades de igualdade, pois ela sinaliza que o relacionamento entre duas classes relacionadas por uma propriedade desse tipo é bidirecional. Assim, quando há o relacionamento dessa classe, verifica-se que a relação pode ocorrer tanto de um sujeito para um objeto, quanto desse mesmo objeto para o sujeito. Para a recuperação da informação, é possível obter as relações que um termo possui, tanto aquelas em que o termo é o sujeito da relação, quanto aquelas em que o termo de busca é o objeto.
<i>owl:AllDisjoint Properties</i>	Diferenciação	Essas propriedades sinalizam que duas propriedades são diferentes. No que tange à recuperação da informação, sinalizar essa diferenciação poderá auxiliar a encontrar o significado de algumas relações existentes, permitindo que a descoberta das relações ocorra com mais eficiência.
<i>owl:propertyDisj ointWith</i>		

Propriedade	Divisão	Características
<i>owl:inverseOf</i>	Inverso	Duas propriedades inversas sinalizam que, caso uma propriedade ligue uma classe com uma segunda, essas duas classes estão relacionadas pela propriedade inversa, e essa relação ocorre com a segunda classe como o sujeito da relação e a primeira classe como o objeto. No contexto da recuperação da informação, propriedades inversas podem ser utilizadas para descobrir novas relações, ao indicar que há um outro relacionamento vinculado que pode auxiliar na descoberta de informações.
<i>owl:AsymmetricProperty</i>	Assimetria	Propriedade assimétrica sinaliza o contrário do que o apontado pela propriedade simétrica, ou seja, uma relação de um sujeito com um objeto não ocorre do objeto para o sujeito. Com efeito inverso do obtido pela propriedade simétrica, para a recuperação da informação, uma relação com essa propriedade sinaliza que a relação de objeto com o sujeito não pode ser extraída para fins de localizar outras relações.

Fonte: Elaborado pelo autor.

O quadro 8 apresenta como as propriedades de propriedades da OWL devem ser utilizadas no contexto da recuperação da informação. Tais elementos serão a base para a identificação de semântica do chamado motor de SPARQL, que será responsável por explorar uma ontologia OWL na busca de relações de um determinado termo.

As características apontadas de cada uma das propriedades da OWL, demonstradas no quadro, serão utilizadas como guia para o motor de SPARQL, ao localizar uma relação contendo aquela determinada propriedade. Dessa forma, os procedimentos tomados pelo motor de SPARQL ao encontrar cada um dos tipos de propriedades estão descritos no quadro 9.

Quadro 9 - Ações das propriedades de propriedades

Divisão	Ação
Hierarquia	Obter a hierarquia de propriedades, verificando se o termo de busca está ligado a algum outro conceito por meio de alguma dessas propriedades.
Relacionamento entre diversos conceitos relacionados	Obter todos os termos que estão ligados ao termo principal, identificando as classes que estão ligadas por meio das propriedades que têm essa característica.
Igualdade	Verificar se as duas ou mais relações que possuem essa característica estão ligando dois conceitos e caso isso ocorra deve excluir-se uma das relações.
Simetria	A identificação de relações simétricas permitirá que o motor de SPARQL possa traçar uma relação de forma bidirecional, ou seja, serão realizadas buscas com um termo tanto ele sendo o sujeito quanto o objeto de uma expressão RDF.
Diferenciação	A identificação de relações diferentes levará à sinalização de que relações de um mesmo sujeito com um predicado ligado por duas propriedades diferentes são diferentes e, portanto, devem ser tratadas individualmente.
Inverso	O motor de SPARQL poderá localizar outras classes na ontologia, ao buscar pelas classes relacionadas por meio da propriedade inversa, inserindo o termo inicial como o objeto da consulta.
Assimetria	Ao localizar uma relação assimétrica, o motor de SPARQL poderá sinalizar que a relação contrária (trocando o sujeito pelo objeto) não deverá ser executada.

Fonte: Elaborado pelo autor.

Embasado no quadro 9, observa-se que a estruturação do motor de SPARQL irá percorrer a ontologia considerando aspectos relevantes da estrutura semântica de um determinado domínio. Em síntese, será capaz de explorar com mais profundidade as relações existentes, permitindo a obtenção de um número maior de termos relacionados ao termo de busca, além de extrair de tais relações uma quantidade maior de características.

Os quadros 7 e 9 (os de ação, tanto do motor de SPARQL, quanto o da expressão de busca) guiarão o desenvolvimento do motor de SPARQL e a montagem da expressão de busca, respectivamente. Essa definição conceitual permitirá que o processo de recuperação da informação utilizando ontologias extraia as principais características semânticas que esse artefato, ontologia OWL, possui.

A camada semântica apresentada nesta subseção possui uma direta relação com a camada de dados, que será abordada na sequência.

5.4 Camada de Dados

A camada de dados é responsável pelas fontes informacionais do modelo, que pode ser considerado um modelo de interoperabilidade de repositórios digitais. No âmbito desta pesquisa, a camada de dados é responsável por duas funções principais, a recuperação dos objetos digitais e o gerenciamento das fontes informacionais.

Tal camada tem como base um modelo de interoperabilidade de repositórios digitais, dando a possibilidade dos usuários interagirem com as fontes informacionais. Essa interação busca fazer com que os usuários possam escolher as fontes de informação (no âmbito da interoperabilidade de repositórios digitais, os provedores de dados) de uma busca. A proposta teórica da camada de dados insere a opção do usuário definir os repositórios digitais que serão utilizados para a realização da busca. Vale destacar que, caso o usuário deseje realizar a busca em algum repositório não cadastrado no sistema, é dada a opção de esse outro repositório ser inserido nele.

A opção de inserir um novo repositório exige que o usuário insira algumas informações sobre essa fonte, para que o sistema realize a coleta (*harvesting*) de seus metadados.

Um dos requisitos para que um repositório seja inserido no sistema é que o usuário esteja logado. Tal processo mostra-se necessário para que exista um registro dos repositórios que as pessoas estão inserindo no sistema, além de permitir que o usuário

tenha possibilidade de configuração e de personalização das fontes informacionais utilizadas em suas buscas. Na sequência serão explicados os processos referentes à camada de dados.

Primeiramente, o usuário interage com a camada de apresentação, como relatado anteriormente, em que esse indivíduo poderá realizar uma busca, escrevendo uma expressão de busca, e poderá selecionar os repositórios digitais que serão as fontes informacionais utilizadas. Além dessa primeira interação, o usuário poderá inserir outros repositórios como fontes de informação; assim, ao usuário agregar uma nova fonte de informação, o modelo deverá inserir esse novo repositório, realizando o *harvesting* dos metadados dele.

Para realizar o processo de *harvesting* deve-se utilizar o protocolo OAI-PMH, que permite que um programa realize a coleta de todos os metadados presentes em um determinado repositório. Deve-se, assim, realizar uma implementação de um programa simples, que realiza a extração dos metadados, utilizando o processo de *harvesting*.

Os metadados extraídos devem ser armazenados em uma base de dados única, em um determinado formato. O uso do OAI-PMH auxilia nessa questão, pois o retorno dos metadados coletados ocorre em padrão *Dublin Core*, auxiliando na existência de um modelo único para a realização da indexação e das buscas, além de auxiliar no modo como as informações serão apresentadas aos usuários.

Vale destacar que uma das questões que deverão ser tratadas pelo modelo é a atualidade dos registros do repositório. Isso porque um repositório digital é atualizado frequentemente, necessitando haver verificações periódicas da existência de novos registros. A indexação computacional acontecerá automaticamente após a coleta de novos metadados, processo esse que apresenta distinta importância, pois quando se trata de repositórios, há um grande volume de metadados que deverão ser indexados e buscados.

Os procedimentos apresentados, tratando da camada de dados, centram-se essencialmente em uma proposta de interoperabilidade de repositórios digitais, que insere como opção a possibilidade do usuário inserir outras fontes de informação durante a realização de uma busca. Tal questão permite a personalização da busca pelos usuários, tornando o processo mais dinâmico e interativo, uma vez que ao se integrar com as demais camadas, especialmente a camada semântica, se permite ao usuário uma gama de opções para aprimorar a sua busca.

Visando validar o modelo proposto foi realizada uma prova de conceito com a implementação de um protótipo dessa proposta. O próximo capítulo mostra como foi realizada tal implementação.

6 IMPLEMENTAÇÃO DO PROTÓTIPO

Na busca de comprovar o modelo proposto nesta pesquisa, foi implementado um protótipo, capaz de executar os principais pontos discutidos no trabalho. Metodologicamente, tal protótipo tem a função de ser uma prova de conceito, em que as principais contribuições da pesquisa são expostas e demonstradas.

A implementação foi desenvolvida por meio da linguagem de programação orientada a objetos Java, pois ela se encontra bastante integrada com as ferramentas da Web Semântica, além de possibilitar a fácil implementação do protocolo OAI-PMH no provedor de serviços.

Como relatado, essa prova de conceito foi construída com o intuito de validar o modelo, visando demonstrar que a proposta construída é factível. O protótipo não tem a função de ser uma implementação formal da proposta, pois essa não é a meta desta pesquisa, mas sim ser um suporte na validação dos objetivos propostos. Um outro destaque para esta seção de implementação é o fato de apontar com mais detalhes a concepção do modelo.

Resumidamente, o protótipo tem duas funções principais, a primeira é permitir que o usuário realize uma busca e a segunda que ele possa cadastrar uma nova fonte informacional. O primeiro processo, relativo à busca, inicia com o usuário acessando uma página de busca, em que ele deverá digitar um termo que representa a sua busca, além de depositar no sistema uma ontologia, que represente o domínio no qual a busca se encontra. O usuário deve ainda, nesse primeiro contato com o sistema, escolher os repositórios em que será realizada a busca dos objetos digitais. Cabe ressaltar que, caso o sistema não tenha cadastrado algum repositório que o usuário deseje, este poderá realizar tal processo.

Posteriormente à definição dessas informações pelo usuário, o sistema deverá processar a ontologia, realizando relacionamentos entre o termo de busca e esse artefato computacional. Esse processo apresentará, como resultados, uma lista contendo uma série de termos relacionados à busca, bem como o significado semântico que tais relações apresentam. Novamente, o usuário irá interagir com o sistema, de modo a escolher, dentre as relações encontradas, aquelas que estão coerentes com suas necessidades informacionais. Essa fase ocorre em uma tela, contendo as relações com seus respectivos significados, em que o usuário deverá marcar somente as relações que ele escolheu.

Posteriormente, o sistema deve processar as escolhas do usuário, de acordo com a ação que cada propriedade da OWL possui dentro da recuperação da informação. Essa

definição é realizada a partir da correlação entre as ações de cada propriedade da OWL na recuperação da informação, apresentada na subseção do “5.3.1 OWL na recuperação da informação”, tendo como consequência uma nova expressão de busca, que será utilizada para localizar, dentro do provedor de serviço, os objetos digitais, que reúnem os repositórios digitais escolhidos pelo usuário.

A última fase desse processo é a apresentação dos resultados obtidos, em que são listados, em um formato tradicional de busca, os resultados que atenderam à expressão de busca construída a partir da interação entre o usuário, o sistema e a ontologia. Nessa listagem, o usuário poderá ver alguns metadados dos registros, além de poder acessar todos os metadados de um registro, que lhe dão acesso ao objeto digital no repositório digital de origem.

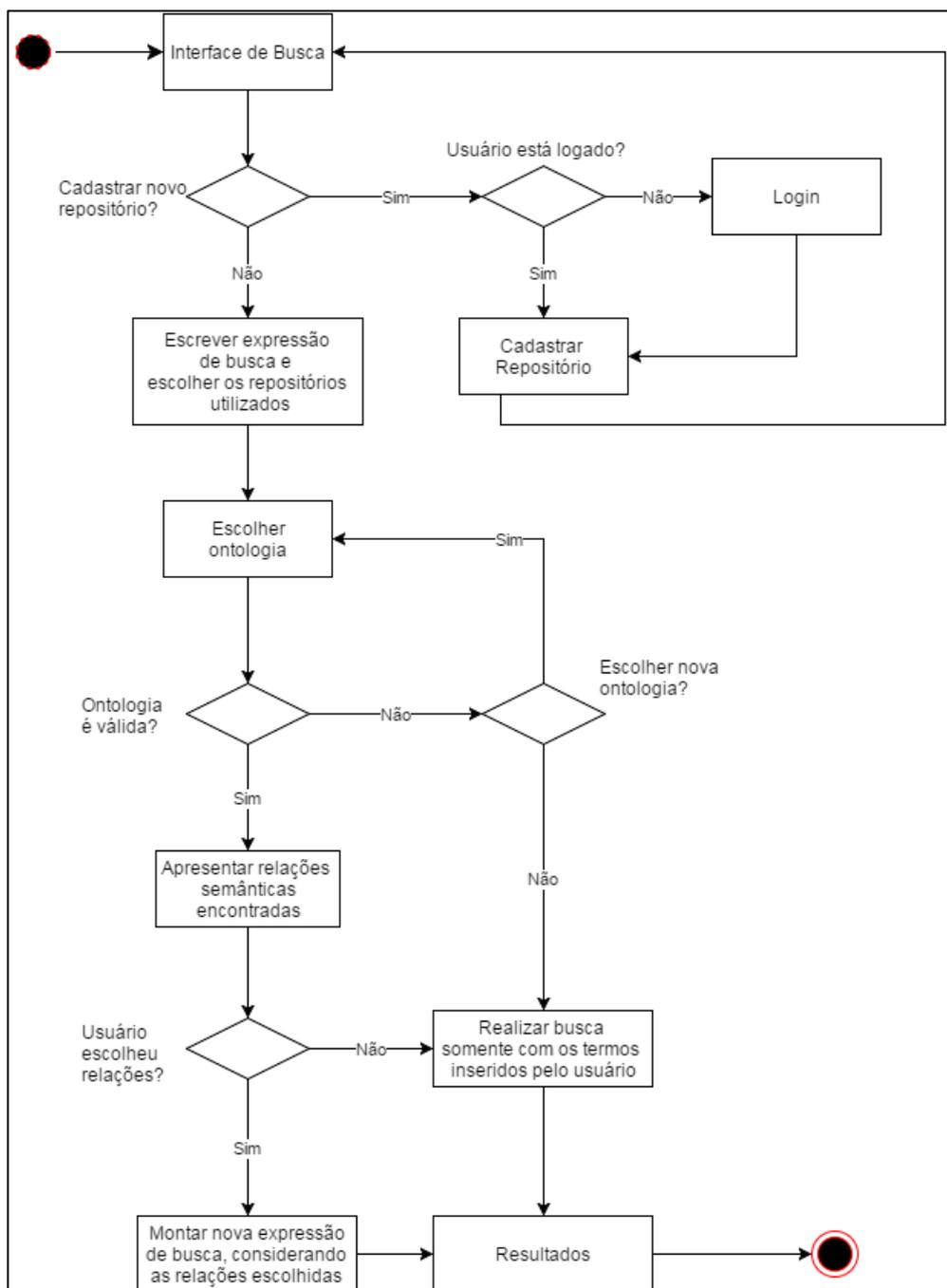
O outro processo que o protótipo possibilita é a inserção de novas fontes informacionais. Essa possibilidade visa dar versatilidade ao modelo, uma vez que dá ao usuário um nível elevado de personalização, ao permitir a escolha e o cadastro de repositórios digitais não listados no sistema.

Esse processo exige que o usuário esteja *logado*, para que seja registrado quem cadastrou os repositórios, além de permitir que o sistema armazene as preferências dos usuários. Dessa forma, com o usuário *logado*, este deverá inserir algumas informações sobre o repositório que será cadastrado, como uma descrição e a URL contendo esse repositório. O sistema, a partir da inserção dessas informações, realizará o processo de *harvesting*, em que coletará os metadados do repositório cadastrado, permitindo que as novas buscas utilizem esse repositório.

Com o intuito de demonstrar tais questões, o primeiro processo realizado para a construção deste protótipo, foi a de um Diagrama de Atividades⁴. Este diagrama de atividades encontra-se ilustrado na Figura 10. Na figura, os losangos representam pontos de decisões, em que o sistema e o usuário deverão escolher uma das opções apresentadas; já os retângulos contemplam atividades realizadas pelo sistema.

⁴ Um diagrama de atividades é um artifício para o desenvolvimento de sistemas computacionais, que visa demonstrar os fluxos executados por um sistema.

Figura 10 - Diagrama de atividades do modelo



Fonte: Elaborada pelo autor.

O diagrama da figura 10 contempla todas as fases de interação entre o protótipo e os usuários, desde a possibilidade de inserção de um novo repositório, passando pelo contato do usuário com as relações semânticas, até a apresentação dos resultados obtidos. Vale destacar que o diagrama de atividades aponta que o protótipo dá a opção de que o usuário não realize uma busca semântica, caso ele não escolha uma ontologia válida, ou não selecione nenhuma das relações semânticas encontradas. Uma ontologia válida pode

ser entendida como uma estrutura computacional bem formada, que pode ser entendido como estrutura que segue as regras de sintaxe de uma linguagem computacional.

No entanto, o fluxo abrange, essencialmente, as questões relativas à participação de ontologias na construção de uma expressão de busca, que considere o contexto, permitindo que o usuário possa escolher as relações semânticas que atendam às suas necessidades informacionais.

Para realizar a implementação desse protótipo, desenvolvemos as quatro camadas descritas no capítulo do modelo (camada de apresentação, camada de controle, camada semântica e camada de dados), seguindo o diagrama de atividades ilustrado na Figura 10. Na sequência será explicitado o modo como ocorreu a implementação dessas quatro camadas.

6.1 Camada de Apresentação

A camada de apresentação é responsável pela interação entre o usuário e o sistema; nela, as diversas telas construídas devem apresentar claramente as suas funções, que impactam diretamente na execução das atividades do protótipo.

A camada de apresentação consiste de seis tipos de telas principais: 1) “Home - busca inicial”, em que o usuário realiza sua pesquisa; 2) “Relações encontradas”, em que são apontadas as relações encontradas na ontologia, com a busca do usuário; 3) “Resultados”, em que são listados todos os resultados encontrados; 4) “Resultado”, em que são apontados os metadados e o *link* do resultado que o usuário encontrou; 5) “Cadastramento de novo repositório”, em que o usuário pode inserir um novo repositório na camada de dados; e 6) “Login”, para que o usuário possa se registrar antes de inserir um repositório no sistema. Todas as telas foram projetadas e desenvolvidas dentro do contexto desta pesquisa.

A primeira tela, demonstrada na Figura 11, contém a barra de pesquisa, em que o usuário pode escrever sua expressão de busca, inserindo os termos que deverão ser recuperados no serviço de repositórios. Além disso, o usuário deverá escolher, na mesma tela, qual será a ontologia a ser utilizada para realizar a expansão semântica e serem encontrados os relacionamentos existentes. A inserção da ontologia poderá ocorrer tanto realizando o *upload* do arquivo, ou inserindo a URL que contém a ontologia. Outra escolha que o usuário deverá realizar diz respeito aos repositórios que serão utilizados na busca, selecionando somente aqueles que serão utilizados como fontes informacionais dela.

Figura 11 - Tela de pesquisa

Interoperabilidade Início

BUSCA EM REPOSITÓRIOS

Escreva sua pesquisa

Ontologia - Arquivo
Escolher arquivo Nenhum arquivo selecionado

Ontologia - URL

Selecionar	Repositório
<input type="checkbox"/>	LUME
<input type="checkbox"/>	UNESP

BUSCAR

unesp GPnti FAPESP ITSP

Fonte: Elaborada pelo autor

A Figura 12 representa as relações existentes entre os termos e a ontologia, na tela chamada de “Relações encontradas”, em que é sinalizado o termo, a relação e um *checkbox*, para o usuário selecionar se deseja utilizar aquele determinado termo em sua busca. Os termos apresentados representam as relações que foram localizadas entre o termo de busca escrito pelo usuário com as outras classes da ontologia. Esse processo, relatado na subseção da camada semântica do modelo, tem como base a descoberta do contexto em que o termo de busca se encontra, sendo solicitado posteriormente que o usuário escolha quais desses termos localizados estão de fato relacionados às suas necessidades informacionais.

Vale destacar que esse processo utiliza as propriedades da OWL para identificar os tipos de relações encontradas, tendo como consequência a apresentação dos termos ao usuário, seguindo a ação que cada propriedade apresenta no âmbito da recuperação da informação. Essas divisões são expressas ao usuário, ao dividir as relações de acordo com cada divisão, o que pode ser visualizado na figura 12 pelo título da divisão, acima dos termos relacionados.

Figura 12 - Tela com as relações encontradas

Interoperabilidade Início

Relações Busca: StEmilion

Aproximação

Termo	Selecionar
Bordeaux	<input type="checkbox"/>
StEmilion	<input type="checkbox"/>
ConsumableThing	<input type="checkbox"/>
CabernetSauvignonGrape	<input type="checkbox"/>
Red	<input type="checkbox"/>
BordeauxRegion	<input type="checkbox"/>
Strong	<input type="checkbox"/>

Hierarquia

Termo	Selecionar
Wine	<input type="checkbox"/>
PotableLiquid	<input type="checkbox"/>
Thing	<input type="checkbox"/>

Diferenciação

Termo	Selecionar
EdibleThing	<input type="checkbox"/>
MealCourse	<input type="checkbox"/>

Igualdade

Termo	Selecionar
StEmilion	<input type="checkbox"/>
Bordeaux	<input type="checkbox"/>
ConsumableThing	<input type="checkbox"/>
Wine	<input type="checkbox"/>
PotableLiquid	<input type="checkbox"/>
Thing	<input type="checkbox"/>
Resource	<input type="checkbox"/>

Fonte: Elaborada pelo autor

Posteriormente à seleção dos termos relacionados, é realizada a montagem da expressão de busca, que será detalhada nas próximas subseções. Utilizando essa nova expressão, é realizada a busca, sendo obtidos os resultados que seguem a lógica construída. A Figura 13 ilustra a tela desses resultados. Vale destacar que os resultados apresentam alguns metadados, um *link* para acessar o registro completo, além da fonte daquele objeto digital, ou seja, o repositório de origem.

Figura 13 - Tela de resultados

Interoperabilidade		Início	Novo Repositório	Fazer Login
Resultados				
Registros				
[Resistência do solo à penetração e ao cisalhamento em diversos usos do solo em áreas de preservação permanente] - LUME				
Autor	Data			
[da Silva, Reginaldo Barboza]	[Wed Feb 29 21:00:00 BRT 2012]			
Palavras-Chave	SCORE			
[Soil quality]	0.07835823			
[Resistência do solo à penetração e ao cisalhamento em diversos usos do solo em áreas de preservação permanente] - UNESP				
Autor	Data			
[da Silva, Reginaldo Barboza]	[Wed Feb 29 21:00:00 BRT 2012]			
Palavras-Chave	SCORE			
[Soil quality]	0.07835823			
[Cadmium availability and accumulation by lettuce and rice] - LUME				
Autor	Data			
[Malavolta, Euripedes]	[Mon Feb 28 21:00:00 BRT 2011]			
Palavras-Chave	SCORE			
[human health]	0.05223882			

Fonte: Elaborada pelo autor

O quarto modelo de tela pode ser visualizado na Figura 14, que contém informações dos metadados de um artigo, além do link para ser acessado o documento PDF, que está armazenado no provedor de dados.

Figura 14 - Tela dos metadados do artigo

Interoperabilidade		Início
Geoprocessing applied to the assessment of environmental noise: a case study in the city of Sorocaba, São Paulo, Brazil		
Autor:	Costa, Samuel Barsanelli Lourenço, Roberto Wagner	
Auxílio:	Universidade Estadual Paulista (UNESP)	
Resumo:	Noise mapping has been used as an instrument for assessment of environmental noise, helping to support decision making on urban planning. In Brazil, urban noise is not yet recognized as a major environmental problem by the government. Besides, cities that have databases to drive acoustic simulations, making use of advanced noise mapping systems, are rare. This study sought an alternative method of noise mapping through the use of geoprocessing, which is feasible for the Brazilian reality and for other developing countries. The area chosen for the study was the central zone of the city of Sorocaba, located in So Paulo State, Brazil. The proposed method was effective in the spatial evaluation of equivalent sound pressure level. The results showed an urban area with high noise levels that exceed the legal standard, posing a threat to the welfare of the population.	
Data:	2014-05-20T13:12:14Z 2014-05-20T13:12:14Z 2011-01-01T00:00:00Z	
Palavras-Chaves:	Environmental noise Noise mapping Geoprocessing Urban planning	
Formato:	329-337	
Cobertura:		
Identificador:	http://dx.doi.org/10.1007/s10661-010-1337-3 Environmental Monitoring and Assessment. Dordrecht: Springer, v. 172, n. 1-4, p. 329-337, 2011. 0167-6369 http://hdl.handle.net/11449/215 10.1007/s10661-010-1337-3 WOS:000284959400024	
Língua:	eng	
Publicador:	Springer	

Fonte: Elaborada pelo autor

As próximas telas apresentadas correspondem à interação que o usuário tem ao inserir novas fontes informacionais.

A tela de cadastro de repositório, ilustrado na Figura 15, indica como um usuário pode inserir um novo repositório na camada de dados, para que esse novo repositório esteja disponível para consulta.

Figura 15 - Tela de cadastro do repositório



Interoperabilidade Início Novo Repositório Fazer Login

Inserir Repositório

Nome Repositório

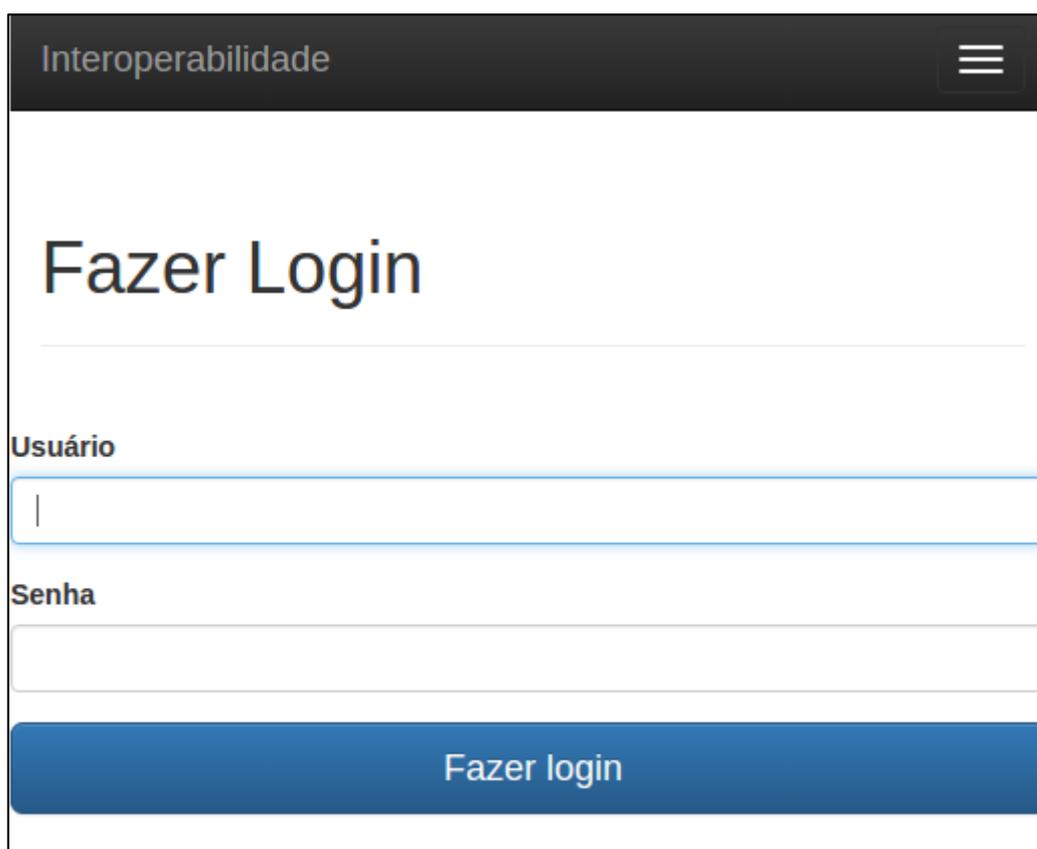
URL Repositório

Inserir Repositório

Fonte: Elaborada pelo autor

A última tela, apresentada na Figura 16, contém as informações para que um usuário possa realizar login no sistema e, assim, possa cadastrar um novo repositório.

Figura 16 - Tela de Login



A imagem mostra a interface de login de um sistema. No topo, há uma barra escura com o texto "Interoperabilidade" à esquerda e um ícone de menu (três linhas horizontais) à direita. Abaixo, o título "Fazer Login" é exibido em uma fonte grande e preta. Seguem dois campos de entrada: "Usuário" e "Senha", cada um com uma borda azulada. Abaixo dos campos, há um botão azul com o texto "Fazer login" em branco.

Fonte: Elaborada pelo autor

Por meio dessas telas, verifica-se que as telas de interação do usuário seguem os processos descritos no diagrama de atividades. Em suma, a camada de visualização é capaz de possibilitar a interação dos usuários com a ontologia, uma vez que os usuários poderão selecionar as relações que atendam às suas necessidades informacionais, além de permitir a interação com a camada de dados, para o usuário conseguir a inserção de novos repositórios no provedor de serviços, assim como a visualização dos resultados das buscas.

A camada de apresentação tem um vínculo direto com a camada de controle, descrita com mais detalhes a seguir.

6.2 Camada de Controle

A camada de controle apresenta funções relativas à definição da ontologia, ao envio dos termos de busca do usuário para a camada semântica e à montagem da nova expressão de busca, sendo um agente computacional que funcionará integrando a camada de apresentação com a camada semântica.

Tal agente apresenta a característica de um serviço Web que recebe e envia solicitações. Para realizar a integração entre o agente controlador e a camada de visualização, foi utilizada a linguagem de programação Java, utilizando o *framework JavaServer Faces* (JSF), que permite que páginas HTML interajam com os códigos construídos dessa linguagem de programação, além de permitir a inserção de códigos nas próprias páginas Web. Assim, quando o usuário realizar uma solicitação, por meio da página de busca, o agente irá realizar a comunicação com as outras camadas, tornando possível que ao final sejam apresentados ao usuário os resultados encontrados.

Para realizar tais processos, o primeiro procedimento consiste na definição da ontologia, em que a camada de controle deve atuar de forma que a ontologia escolhida pelo usuário seja inserida na camada semântica. Esse processo ocorrerá por meio de requisições, que enviam a ontologia, em OWL ou RDF, para a camada semântica.

O procedimento subsequente à definição da ontologia é o envio do termo de busca do usuário para a camada semântica. Esse processo ocorre, novamente, pelo envio de uma requisição à camada semântica, contendo o termo escrito pelo usuário.

Posteriormente ao envio do termo para essa camada, são apresentadas aos usuários, na camada de apresentação, as relações vinculadas com os termos de buscas. A partir disso, o usuário seleciona as relações, de acordo com as suas necessidades informacionais.

Dessa forma, ocorre o último processo da camada de controle, que diz respeito à montagem da nova expressão de busca, que deverá refletir as escolhas dos usuários.

O processo de montagem da expressão de busca ocorrerá de acordo com as relações semânticas escolhidas pelo usuário, estando de acordo com a definição de como as propriedades da OWL atuarão na recuperação da informação, o que está descrito na seção 5.3.1.

A montagem da expressão de busca identifica, primeiramente, todos os termos escolhidos e as suas relações, diferenciando-as. Posteriormente, verifica-se qual é a ação de cada relação. Para esse processo, realizou-se a implementação de uma lista que contém a ação, de acordo com cada tipo de relação.

Conforme apresentado, as ações estão vinculadas a axiomas booleanos, contendo as definições, de acordo com os operadores booleanos, tais como o AND, OR, NOT. A expressão de busca construída seguirá esses operadores.

Tais expressões se embasaram no quadro 7, apresentados na subseção da OWL na recuperação da informação. O processo ocorre ligado diretamente com a camada

semântica, que será responsável por obter os termos relacionados ao termo de busca, bem como os tipos de relações que ligam esses termos.

Vale destacar que serão utilizadas as chamadas propriedades de classes da OWL, pois são elas que fornecem as características semânticas contextuais de cada classe relacionada ao termo de busca. As propriedades de propriedades serão utilizadas como um suporte a esse processo, tornando o processo de descoberta de classes mais semântico, ao considerar o contexto de cada relação. Esse tipo de propriedade será abordado com maiores detalhes na subseção seguinte, que explicita o processo de implementação da camada semântica.

Em síntese, a camada de controle, ao receber os termos relacionados ao termo de busca que foram escolhidos pelo usuário, montará uma nova expressão de busca que refletirá o contexto que cada termo possui, seguindo as características semânticas obtidas da ontologia. Um outro ponto de destaque para a implementação desse processo na camada de controle trata da questão da ordem em que cada termo será inserido na expressão de busca. Tal ordenação foi definida previamente também no quadro 7, em que, a partir das características dos operadores booleanos, foi definida uma ordem a ser seguida.

A implementação desse processo ocorreu por meio de funções que foram criando a expressão de busca de acordo com a lista de relações escolhidas pelo usuário. O primeiro processo consiste em verificar as relações escolhidas pelo usuário, identificando os termos que deverão compor a nova expressão de busca.

Na figura 17 é possível identificar o termo de busca com as suas respectivas relações escolhidas pelo usuário, a partir da ontologia. Nessa figura é possível verificar que está descrito o tipo de relação existente entre o termo de busca e o termo da relação.

Figura 17 - Relações escolhidas entre o termo de busca e a ontologia

```

Termo Inicial: Linked Data
Relação 1
Termo: Dados Ligados
Relacao: owl:equivalentClass
Tipo Relação: igualdade
Relação 2
Termo: Web Semântica
Relacao: rdfs:range
Tipo Relação: aproximacao
Relação 3
Termo: Arquitetura da Informação
Relacao: owl:disjointWith
Tipo Relação: diferenciacao
  
```

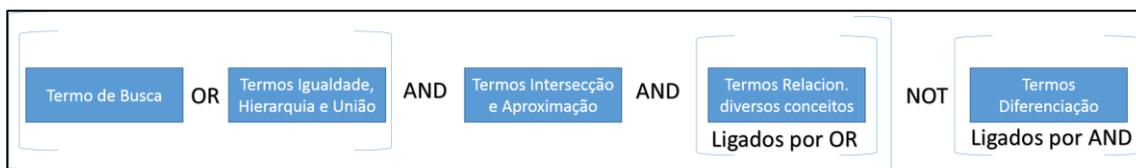
Fonte: Elaborada pelo autor

Embasado nas relações demonstradas na Figura 17, é possível obter a ação que cada uma dessas relações apresenta para a recuperação da informação, como relatado anteriormente. Desta feita, o protótipo inicia um processo em que são realizadas as inserções dos termos escolhidos, relacionados com o termo de busca, seguindo as ações e a ordem devida de cada propriedade da OWL.

O algoritmo construído irá comparar cada propriedade, individualmente, com uma lista que contém a ação e a ordem para a inserção desse termo, colocando tais informações em uma nova lista. Posteriormente, tendo essa lista de termos, ações e ordem completa, inicia-se o processo de montagem da expressão de busca em si. Esse processo começa por ordenar todos os termos de acordo com a ordem de introdução na expressão, para que as relações de um mesmo nível sejam inseridas em subconjuntos da expressão.

Por meio desse ordenamento é possível traçar os relacionamentos de cada termo dentro da expressão. Esse processo segue a ordem de inserção dos termos, que pode ser traduzido na figura 18, em que são demonstradas as formas como os termos são introduzidos na expressão de busca.

Figura 18 - Lógica de inserção dos termos na expressão de busca



Fonte: Elaborado pelo autor

Na figura 18, é possível identificar que cada parte da expressão contém o tipo de relação que cada termo mantém no relacionamento com o termo de busca, de modo que as relações booleanas: OR, AND e NOT são construídas na expressão, levando em conta a ação que cada propriedade possui, sem que existam problemas lógicos na montagem da expressão.

Visando demonstrar um exemplo da construção dessa expressão, a figura 19 contempla um exemplo da expressão de busca, de acordo com as relações escolhidas pelo usuário na ontologia, tendo como lógica de montagem as ações de cada propriedade. A figura 19 representa uma expressão de busca construída baseada no termo “Linked Data”, em que foram selecionadas as relações: 1) do tipo igualdade, termo: dados ligados; 2) do tipo aproximação, termo: Web Semântica; 3) do tipo diferenciação, termo: arquitetura da informação.

Figura 19 - Exemplo de expressão de busca utilizando o termo *Linked Data*

```
((Linked Data OR Dados Ligados) AND Web Semântica) NOT (Arquitetura da Informação)
```

Fonte: Elaborada pelo autor

Com a figura 19, é possível verificar que a nova expressão de busca segue as particularidades do domínio, obtidas com a utilização da ontologia, além de considerar o contexto do usuário, uma vez que ele é o responsável por selecionar somente as relações que estão de acordo com as suas necessidades informacionais.

A camada com mais relação com o controle é a camada semântica, responsável por obter as relações da ontologia, analisando as propriedades existentes, possibilitando que a construção da expressão de busca possa ocorrer. Os detalhes acerca da implementação da camada semântica são descritos na sequência.

6.3 Camada Semântica

A implementação da camada semântica tem o propósito de possibilitar a descoberta de informações da ontologia, permitindo que a camada de controle seja capaz de criar as novas expressões de busca.

Para tal, o primeiro processo desenvolvido tratou da forma como as ontologias são inseridas pelos usuários. Esse desenvolvimento possibilitou que a inserção de ontologias ocorresse, tanto via URL, em que o usuário insere o endereço em que a ontologia se encontra hospedada, fazendo com que o programa realize tanto o *download* do arquivo, quanto a inserção do arquivo da ontologia que está localmente na máquina do usuário.

A partir do instante em que a ontologia é inserida, o protótipo necessita realizar o seu mapeamento. Esse processo ocorre por meio da utilização de algoritmos que convertem a ontologia em uma classe Java chamada de “*OntModel*”, que possibilita a manipulação da ontologia pelos algoritmos, em linguagem de programação, inclusive para a realização de consultas SPARQL, que se dará por meio do programa construído em Java.

Vale destacar que as buscas realizadas nessa fase ocorrem integralmente dentro da ontologia escolhida pelo usuário. No exemplo demonstrado aqui, essa busca ocorreu na ontologia de exemplo recomendada pela W3C para testes⁵.

Na etapa seguinte ocorre a localização do termo na ontologia. Para ocorrer tal processo, é necessário que seja feita uma consulta dinâmica SPARQL, que recupera o termo dentro das relações do RDF, mais especificamente, localiza dentro da posição objeto (no contexto da tripla RDF, sujeito, predicado e objeto).

A localização dentro da ontologia ocorre com a função SPARQL “*regex*”, que irá buscar dentro de cadeias de caracteres, e “*filter*” que irá realizar filtros dentro da busca. A Figura 20 mostra um exemplo de consulta gerada dinamicamente para o termo “StEmilion”, dentro de uma ontologia. Outro detalhe na consulta apresentada é o prefixo “*xsd*”, que é utilizado como prefixo para o tipo de dados “*string*” na função “*BIND*”, sendo que tal função é utilizada para a definição de uma nova variável que será usada dentro do filtro.

⁵ Wine Ontology W3C. Disponível em: <<https://www.w3.org/TR/owl-guide/wine.rdf>>. Acesso em: 25 fev. 2017.

Figura 20 - Função SPARQL de localizar termo

```

PREFIX xsd: http://www.w3.org/2001/XMLSchema#
SELECT DISTINCT ?valor
WHERE {
  ?valor ?qualquer ?description.
  BIND (xsd:string(?valor) as ?string1)
  FILTER (regex(?string1, "StEmilion", "i"))
}

```

Fonte: Elaborada pelo autor

O retorno da consulta apresentada pela Figura 20 será uma lista de resultados que satisfazem tal consulta SPARQL. No exemplo realizado, a consulta retornará o resultado com a seguinte URI: “<http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#StEmilion>”, obtendo, assim, o nome da classe que indica o termo buscado pelo usuário.

Com tal informação localizada, o sistema necessita encontrar todas as relações existentes do objeto com os outros termos da ontologia. Para fazer tal processo, o gerador de consultas SPARQL coloca como sujeito da tripla RDF o objeto encontrado no passo anterior (no exemplo, a URI com o final ‘StEmilion’), armazenando em uma lista todos os predicados e os objetos retornados da consulta. A consulta inicial que o gerador cria contém uma única condição dentro da cláusula “*WHERE*”. A Figura 21 mostra um exemplo de consulta, que utiliza a URI localizada no exemplo anterior.

Figura 21 - Função SPARQL localizar relações

```

PREFIX xsd: http://www.w3.org/2001/XMLSchema#
SELECT DISTINCT ?predicado ?description
WHERE {
  <http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#StEmilion>
  ?predicado ?description.
}

```

Fonte: Elaborada pelo autor

Os resultados obtidos a partir da consulta SPARQL, demonstrado na Figura 21, não conseguem recuperar todas as relações que um termo possui em uma ontologia, devido a existência de nós em branco, o que é melhor relatado no subcapítulo “Módulo de semântica”, no capítulo “Modelo”. Um exemplo para demonstrar a questão envolvendo os nós em branco é dado na Figura 9, em que, para recuperar todos os dados

do endereço do sujeito “José da Silva”, no esquema de geração automática de consultas SPARQL, é necessário que sejam feitas duas consultas, uma para localizar o nó em branco, e uma segunda que irá localizar os nós ligados ao nó em branco. Porém, caso fosse feita uma única consulta, localizando só os nós ligados diretamente ao sujeito, ocorreria uma grande perda de informações, pois só seriam localizados os nós em branco.

A solução para tal questão é a construção de consultas dinâmicas que vão avançando nível a nível desses nós em branco, até que sejam extraídos todos os dados. Retornando ao contexto da descoberta das relações que um termo possui na ontologia, o mesmo processo é fundamental, necessitando que sejam realizadas novas consultas que vão avançando nos nós em branco existentes, para que, assim, possam ser extraídas todas as relações de um termo. A Figura 22 detalha uma consulta gerada, que vai avançando pelos diversos nós em branco, relacionados ao termo principal.

Figura 22 - Consulta SPARQL com nós em branco

```

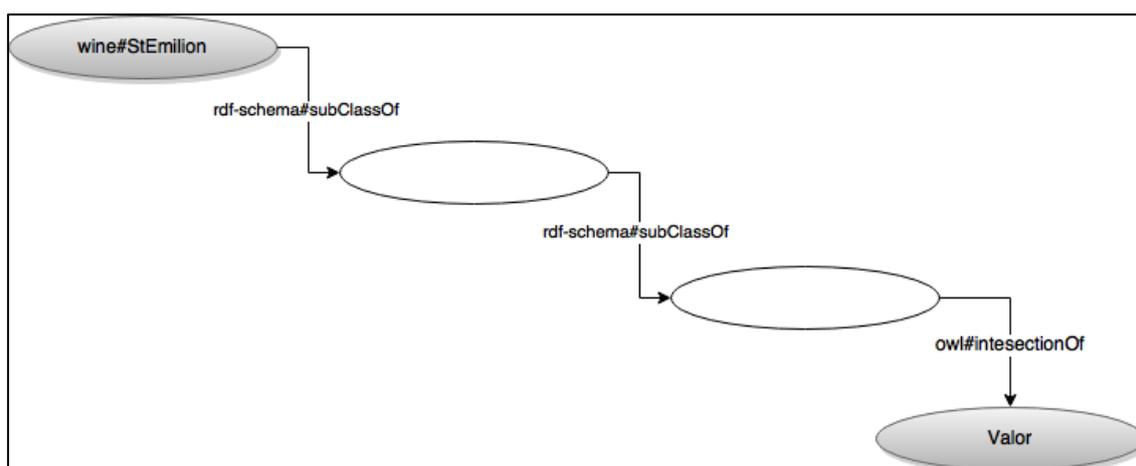
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
SELECT DISTINCT ?predicado ?description
WHERE {
  <http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#StEmilion>
  <http://www.w3.org/2000/01/rdf-schema#subClassOf> ?outradescription1.
  ?outradescription1 <http://www.w3.org/2000/01/rdf-schema#subClassOf>
  ?outradescription2.
  ?outradescription2 <http://www.w3.org/2002/07/owl#intersectionOf>
  ?outradescription3.
  ?outradescription3 ?predicado ?description.
}

```

Fonte: Elaborada pelo autor

Na figura 22 é possível visualizar como o processo ocorre, havendo um nó em branco ligado ao “wine#StEmilion”, representado pela variável “?outradescription1”, também existe um outro nó em branco representado pelo “?outradescription2”, ligado com o nó em branco “?outradescription1”. Por último existe um nó com conteúdo (“?outradescription3”) ligado ao nó em branco ?outradescription2. A Figura 23 ilustra, em esquema de grafo, os dados recuperados por uma consulta SPARQL, com diversos nós em branco.

Figura 23 - Múltiplos nós brancos dentro de uma ontologia representada em grafo



Fonte: Elaborada pelo autor

A consulta demonstrada na Figura 22 contém diversas relações; no entanto, para permitir a geração de uma consulta que avança por diversos nós em branco, é necessário que seja gerada uma nova consulta, a cada vez que for encontrado um novo nó em branco. Como a ontologia é desconhecida do sistema, não há ciência nem da quantidade, nem do local em que serão encontrados os nós em branco; assim, o tratamento deverá ocorrer dinamicamente, sendo utilizados algoritmos e técnicas recursivas para que a cada momento em que for encontrado um novo nó em branco, sejam geradas novas consultas, até serem verificados todos os dados que estão relacionados a um termo na ontologia.

A recursividade é um termo utilizado na Ciência da Computação para falar de uma rotina que vai se repetindo, ao chamar ela mesma, executando as mesmas funções diversas vezes, até atingir uma condição de parada. No caso deste trabalho, a condição de parada é encontrar um nó que contenha valor.

O processo como um todo gera um grande número de consultas, que vai analisando e verificando todos os termos relacionados ao termo principal. Dessa forma, todas as relações são inseridas em uma lista que armazenará a relação e o seu tipo.

Um segundo processo de relevância para esta pesquisa leva em consideração as propriedades da OWL que tratam das propriedades da ontologia. Como relatado na subseção da OWL na recuperação da informação, o motor de SPARQL deverá considerar o contexto em que as propriedades estão contidas, para que, assim, a descoberta da informação seja mais eficaz. O motor de SPARQL deverá, ao encontrar uma relação que apresente alguma das propriedades relatadas, realizar novas consultas explorando outras relações existentes.

Visando demonstrar como as ações indicadas no quadro 9 serão expressas dentro do SPARQL, construíram-se consultas genéricas que apontam a ação tomada em cada um dos casos expressos. O quadro 10 demonstra cada caso, expressando a propriedade em questão com o valor P. Vale destacar que essas consultas não apresentam uma codificação real da consulta SPARQL, mas, sim, uma demonstração do comportamento de cada consulta no âmbito da recuperação da informação. Posteriormente à apresentação dessas consultas, apresenta-se uma implementação real de um exemplo de consulta, funcionando integrada a uma ontologia.

Quadro 10 - Consulta SPARQL genérica para cada divisão de propriedade

Divisão	Consulta SPARQL	Comentário
Hierarquia	<pre>SELECT WHERE{ ?termo_busca PROPRIEDADE1 ?outro_termo1. PROPRIEDADE1 PROP_HIERARQUIA PROPRIEDADE2. ?termo_busca PROPRIEDADE2 ?outro_termo2. }</pre>	Obter a hierarquia de propriedades, verificando se o termo de busca está ligado a algum outro conceito por meio de alguma dessas propriedades mais específicas.
Relacionam ento entre diversos conceitos relacionados	<pre>SELECT WHERE{ ?termo_busca PROPRIEDADE1 ?outro_termo1. ?outro_termo1 PROPRIEDADE1 ?outro_termo2. }</pre>	Tal processo é repetido, até obter todas as propriedades ligadas, localizando outros termos relacionados.
Igualdade	<pre>SELECT WHERE{ ?termo_busca PROPRIEDADE1 ?outro_termo1.</pre>	São localizadas relações iguais, sendo que esta propriedade auxilia para excluir tais relações,

Divisão	Consulta SPARQL	Comentário
	PROPRIEDADE1 PROP_IGUALDADE PROPRIEDADE2. ?termo_busca PROPRIEDADE2 ?outro_termo1. }	deixando somente uma ocorrência do termo
Simetria	<i>SELECT</i> <i>WHERE</i> { ?termo_busca PROPRIEDADE1 ?outro_termo1. ?outro_termo2 PROPRIEDADE1 ?termo_busca. }	Essas relações permitem a descoberta de novos termos relacionados, uma vez que será possível identificar outros termos relacionados por meio da relação simétrica.
Diferenciação	<i>SELECT</i> <i>WHERE</i> { ?termo_busca PROPRIEDADE1 ?outro_termo1. ?termo_busca PROPRIEDADE2 ?outro_termo1. }	A descoberta de que as propriedades 1 e 2 são diferentes poderá ajudar o sistema em considerar que essas duas relações, apesar de possuírem os mesmos sujeito e predicado, serão consideradas duas relações distintas.
Inverso	<i>SELECT</i> <i>WHERE</i> { ?termo_busca PROPRIEDADE1 ?outro_termo1. PROPRIEDADE1 PROP_INVERSAS PROPRIEDADE2. ?outro_termo2 PROPRIEDADE2 ?termo_busca. }	O motor de SPARQL poderá localizar outras classes na ontologia, ao buscar pelas classes relacionadas por meio da propriedade inversa, inserindo o termo inicial como o objeto da consulta, e utilizando a propriedade

Divisão	Consulta SPARQL	Comentário
		inversa no predicado da consulta.
Assimetria	<pre>SELECT WHERE{ ?termo_busca PROPRIEDADE1 ?outro_termo1.</pre>	A partir da identificação de uma relação com propriedade assimétrica, não se deve realizar uma busca invertendo o sujeito pelo objeto, pois essa relação não tem comportamento simétrico.

Fonte: Elaborada pelo autor

Por meio do quadro 10, observa-se que ao se obterem as classes relacionadas a um termo de busca, com suas respectivas propriedades, é possível realizar novas consultas dentro da ontologia, melhorando a recuperação dos termos relacionados. Esse processo ocorrerá após o de localização de termos, em que as consultas SPARQL poderão expandir os resultados obtidos, ao executar as consultas de acordo com as propriedades da OWL.

Visando demonstrar a efetividade desse processo, apresenta-se um exemplo das consultas SPARQL geradas, localizando novas relações, ou inserindo novas características às relações existentes. O exemplo apresenta a consulta do tipo de simetria, que, como relatado, possibilita inverter a ordem do sujeito com o objeto, uma vez que a semântica da relação se mantém.

Desta feita, o motor de SPARQL executa uma consulta em que a busca se dá pelas relações simétricas, localizando diversos relacionamentos existentes. A figura 24 expressa a consulta que o motor de SPARQL gera para considerar as relações simétricas.

Figura 24 - Consulta de simetria gerada

```

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
select DISTINCT ?propriedade ?termo1 ?termo2
where { {
<http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#StEmilion> ?propriedade
?termo1. }
UNION
{?termo2 ?propriedade <http://www.w3.org/TR/2003/PR-owl-guide-
20031209/wine#StEmilion>.
?propriedade rdf:type ?teste FILTER(?teste IN (owl:SymmetricProperty)).} }

```

Fonte: Elaborada pelos autores.

Na figura 24 é possível verificar que são localizadas todas as relações que a entidade possui; além disso, é realizada uma união com os resultados obtidos da inversão de sujeito e objeto, caso a propriedade analisada seja simétrica. Esse comportamento insere a realização de inferências para o motor de SPARQL, uma vez que esse artifício levará em consideração características semânticas da ontologia. A inferência ocorre por expandir a busca, caso a propriedade seja simétrica.

O exemplo apresentado permite a localização de novas relações que não foram localizadas, possibilitando inclusive a realização dos procedimentos descritos dos nós em branco, nas novas propriedades encontradas. Esses procedimentos permitem a descoberta de um número maior de relações, que possuem, de acordo com a ontologia, uma relação direta com a busca realizada pelo usuário.

Após a descoberta de todas as relações existentes, é realizada uma classificação de todos os resultados que foram encontrados, sendo feito uma ordenação, que agrupa pelo tipo de relação. A Figura 25 demonstra um exemplo, contendo um fragmento da lista de relações encontradas a partir das consultas SPARQL realizadas.

Figura 25 - Fragmento das relações existentes em uma consulta

Termo: <http://www.w3.org/TR/2003/PR-owl-guide-20031209/food#ConsumableThing>	
Relação: <http://www.w3.org/2000/01/rdf-schema#subClassOf>	
Termo: <http://www.w3.org/2002/07/owl#Thing>	Relação:
<http://www.w3.org/2000/01/rdf-schema#subClassOf>	
Termo: <http://www.w3.org/TR/2003/PR-owl-guide-20031209/food#ConsumableThing>	
Relação: <http://www.w3.org/2000/01/rdf-schema#subClassOf>	
Termo: <http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#StEmilion>	Relação:
<http://www.w3.org/2002/07/owl#equivalentClass>	
Termo: <http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#Bordeaux>	Relacao:
<http://www.w3.org/2002/07/owl#equivalentClass>	

Fonte: Elaborada pelo autor

Assim, estando a lista organizada, a parte de interface visual do protótipo deverá utilizar tais relações para que os usuários possam selecionar as que serão utilizadas para auxiliar na recuperação de informação. Essa escolha embasará o processo detalhado na subseção 6.2, que explicita os procedimentos da camada de controle.

Em suma, a camada semântica recupera todas as expressões relacionadas, bem como as propriedades que caracterizam cada uma dessas relações, ficando a cargo da camada de controle criar a nova expressão de busca, que considerará as propriedades da OWL relacionadas a cada termo.

Em um nível mais baixo, os processos demonstrados nas camadas de apresentação, de controle e de semântica estarão relacionados à camada de dados, que é responsável pela interação do sistema com os dados colhidos dos repositórios digitais. A implementação dessa camada está detalhada na próxima subseção.

6.4 Camada de Dados

O protótipo construído neste trabalho tem como base um provedor de serviços, que permite o acesso a diversos repositórios digitais, nesse contexto chamados de provedores de dados. Vale destacar que o processo de inserção de repositórios ocorre dinamicamente, de forma que o usuário poderá escolher uma nova fonte informacional para ser inserida no sistema.

Em síntese, a camada de dados realiza dois processos: a inserção dos registros de metadados e o processo de recuperação da expressão de busca. Para a execução dos processos foram implementados dois serviços que interagem: 1) “Servidor Solrj”, responsável pela interação entre o restante do sistema e o Apache SOLR, em que o

“Servidor Solrj” tem esse nome por utilizar como principal ferramenta um conjunto de códigos programados com a Linguagem Java, construído com o intuito de fazer a comunicação entre essa linguagem e o Apache Solr; e 2) “Servidor Solr”, responsável pela indexação computacional, armazenamento e busca dos metadados dos repositórios digitais, extraídos por meio do *harvesting*.

Os dois processos supracitados realizados por essa camada serão abordados a seguir.

6.4.1 Coleta e Armazenamento dos Metadados

A implementação do esquema de interoperabilidade em repositórios digitais seguiu o modelo descrito no capítulo “Modelo”, em que são identificados os pontos necessários para a implementação.

O primeiro processo é o *harvesting*, em que é executado a coleta dos metadados de distintos repositórios. Esse processo ocorre tanto periodicamente, para atualizar os registros dos repositórios já cadastrados, quanto a cada vez que um usuário deseja inserir um novo repositório para serem realizadas buscas.

No âmbito deste trabalho, realizou-se a coleta no repositório institucional UNESP, que contém a produção intelectual da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP).

A conexão ao serviço OAI-PMH e a solicitação dos metadados ao servidor do repositório ocorre por meio da utilização de uma URL que contém as informações desejadas. Tal URL engloba, em suma, o sítio onde está hospedado o repositório, a indicação que se refere ao serviço de OAI-PMH e a requisição em si. Cabe destacar que o usuário, ao solicitar o cadastramento de um novo repositório, deverá inserir somente a sua URL, como por exemplo “<http://repositorio.unesp.br>” ou “<http://www.lume.ufrgs.br>”, e o sistema criará uma nova URL contendo as informações necessárias. Assim, nos exemplos relatados, as URLs ficariam:

“http://repositorio.unesp.br/oai/request?verb=ListRecords&metadataPrefix=oai_dc” e
“http://www.lume.ufrgs.br/oai/request?verb=ListRecords&metadataPrefix=oai_dc”.

Recebendo a requisição, o provedor de dados retorna uma lista com os metadados, que atende à solicitação feita pelo agente computacional, utilizando como sintaxe linguagem XML. Como relatado no capítulo do protocolo OAI-PMH, esse protocolo utiliza como padrão de metadados o *Dublin Core*. Um fragmento de um registro retornado pode ser visualizado na Figura 26.

Figura 26 - Fragmento do registro retornado pelo provedor de dados

```

<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" >
<dc:title>Effects of morin on snake venom phospholipase A(2) (PLA(2))</dc:title>
<dc:creator>Iglesias, C. V.</dc:creator>
<dc:creator>Aparicio, R.</dc:creator>
<dc:creator>Rodrigues-Simioni, L.</dc:creator>
<dc:creator>Camargo, E. A.</dc:creator>
<dc:creator>Antunes, E.</dc:creator>
<dc:creator>Marangoni, S.</dc:creator>
<dc:creator>Toyoma, D. D.</dc:creator>
<dc:creator>Toyama, M. H.</dc:creator>
<dc:contributor>Universidade Estadual Paulista (UNESP)</dc:contributor>
<dc:subject>PLA(2)</dc:subject>

```

Fonte: Elaborada pelo autor

Contudo a resposta fornecida pelo provedor de dados não é compatível com a ferramenta de indexação, que será melhor explorada na sequência. Dessa forma, é necessário que seja realizado uma conversão do XML retornado pelo servidor do repositório digital, para o formato exigido na ferramenta de indexação. A Figura 27 mostra o registro, ilustrado na Figura 26, após ser executado o processo de conversão.

Figura 27 - Fragmento de um registro XML

```

<doc>
<field name="title">Kinetics and Adsorption Isotherms of Bisphenol A, Estrone, 17 beta-
Estradiol, and 17 alpha-Ethinylestradiol in Tropical Sediment Samples</field>
<field name="creator">Cunha, Bruno B.</field>
<field name="creator">Botero, Wander Gustavo</field>
<field name="creator">Oliveira, Luciana Camargo</field>
<field name="creator">Carlos, Viviane M.</field>
<field name="creator">Pompeo, Marcelo L. M.</field>
<field name="creator">Fraceto, Leonardo F.</field>
<field name="creator">Rosa, André Henrique</field>
<field name="contributor">Universidade Estadual Paulista (UNESP)</field>
<field name="subject">Endocrine disruptors</field>
<field name="subject">Sediment</field>
</doc>

```

Fonte: Elaborada pelo autor

A Figura 27 demonstra a estrutura utilizada para serem inseridos arquivos na ferramenta de indexação. Para que um registro seja inserido, ele deve estar contido dentro de uma *tag* “doc”, e os campos pertencentes a tal registro devem estar necessariamente especificados na *tag* “field”, tendo seu nome descrito no qualificador “name”.

Posteriormente, tratou-se da questão da indexação computacional dos metadados. Para realizar tal etapa, buscou-se por ferramentas de indexação computacional automática, que apresentassem uma eficiência na indexação e na busca dos registros. Esse processo levou em consideração a capacidade das ferramentas de atenderem rápida e eficazmente às necessidades de informação dos usuários.

O software que melhor atendeu a tais requisitos foi o Apache Solr. Tal ferramenta possibilita que seja realizada a indexação de um grande número de arquivos, permitindo uma posterior recuperação. Destaca-se, no Apache Solr, a velocidade com que a indexação e as buscas são realizadas, além de ele ser baseado no Apache Lucene, que abre uma gama de possibilidades para realizar a recuperação de informação, no que diz respeito à criação de expressões de busca

No âmbito desta pesquisa, o Apache Solr foi configurado para que a indexação dos documentos acontecesse baseada nos quinze elementos descritores do *Dublin Core*. Outra característica da ferramenta, nesta pesquisa, foi que os arquivos incluídos no sistema de indexação apresentavam o formato XML. A Figura 27 apresenta a sintaxe padrão para realizar a indexação no Apache Solr.

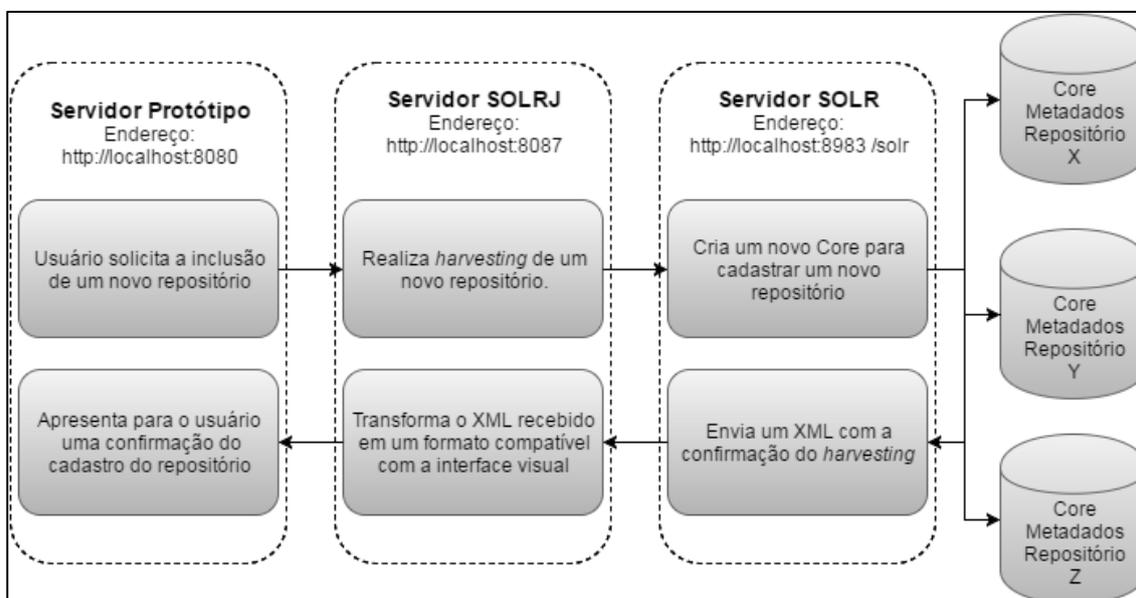
Para utilizar essa API foi construído um servidor, que recebe solicitações, tanto para realizar a indexação, quanto para a realização de buscas, realizando, basicamente, essas duas funções: indexação, em que o servidor recebe um arquivo XML com os documentos para serem indexados, e busca, que será explorada na próxima subseção. Esse servidor foi construído seguindo o padrão de *sockets*, funcionando em esquema de cliente servidor, em que o cliente solicita, por meio de um XML, alguma informação, obtendo como retorno um outro XML, com a informação desejada.

O processo de indexação computacional ocorre com uma solicitação do agente controlador, que irá enviar periodicamente requisições para o servidor do SOLRJ, contendo a URL do provedor de serviços onde deverá ser feita a coleta dos metadados. Obtendo tal URL, o servidor SOLRJ irá realizar o processo de coleta de metadados, obtendo, assim, os metadados em um formato que possa ser indexado. A partir disso, será realizada uma transformação do XML, em classes chamadas *SolrInputDocument*, contendo todas as informações do XML. Como são extraídos diversos registros do processo de coleta de metadados, é utilizada uma lista contendo diversos *SolrInputDocument*, para, assim, serem feitas as inserções no servidor Apache SOLR, e como consequência, serem incluídos e indexados na base de dados do Apache SOLR.

Destaca-se, ainda, que o sistema irá criar um *core* para cada repositório inserido, dividindo os registros, para que ao se realizar uma busca ela seja feita somente nos repositórios definidos pelo usuário. Esses *cores* são divisões dentro da base de dados, criando ambientes separados para a inserção dos dados.

Sintetizando o processo relatado, a Figura 28 apresenta o modo como ocorrem as relações entre os diversos serviços necessários para o funcionamento da coleta e do armazenamento dos metadados dos repositórios. Nessa figura são apresentadas, ainda, informações acerca das URLs em que os serviços funcionaram durante os testes, que ocorreram em ambiente de desenvolvimento local.

Figura 28 - Síntese do processo de coleta e de armazenamento dos metadados



Fonte: Elaborada pelo autor.

Por meio da figura 28, é possível identificar as relações entre os serviços, necessários para a coleta e o armazenamento dos metadados dos diversos repositórios. Além disso, verifica-se que o usuário pode inserir um repositório, que se tornará um novo core de metadados, representados como os cilindros na parte direita da figura.

O processo de busca está diretamente relacionado às questões expostas sobre a coleta e o armazenamento dos metadados. Na próxima seção será explorada essa questão, identificando as relações existentes entre essas duas etapas.

6.4.2 Processo de Recuperação da Informação no Provedor de Serviços

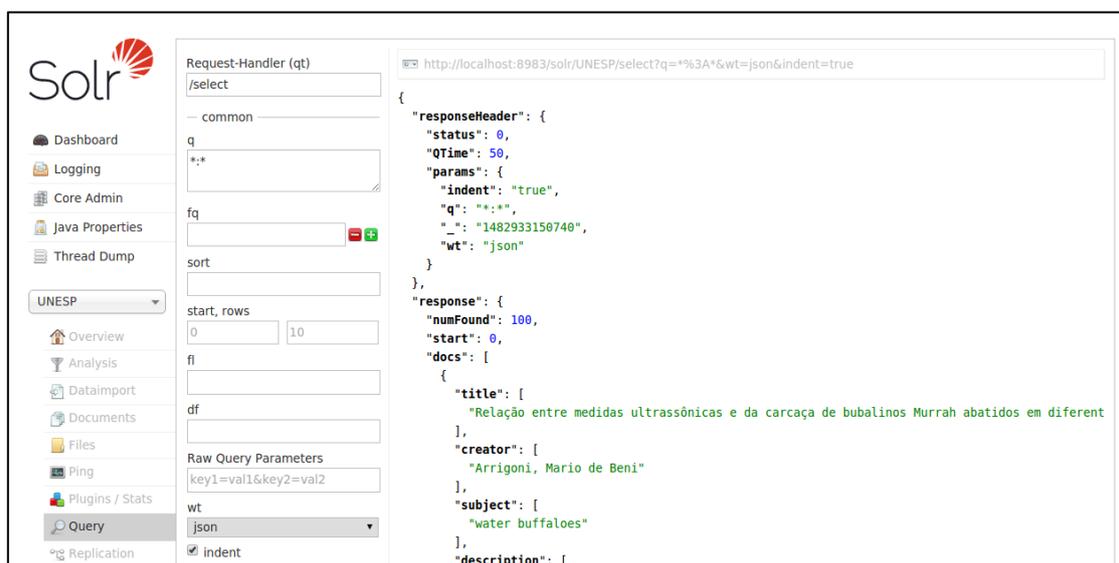
A recuperação da informação dentro dos registros coletados nos repositórios digitais ocorre por meio da expressão de busca que reflete as necessidades informacionais dos usuários, que já está modificada de acordo com as interações entre os usuários e a camada semântica.

Dessa forma, o processo de busca irá utilizar-se dos serviços: servidor Solrj e servidor Solr, que foram citados na subseção anterior, pelo papel deles nos processos de coleta e de armazenamento dos metadados.

Para discorrer sobre os processos executados nesse trabalho, é necessário antes apontar as formas de execução das buscas dentro do Apache Solr. Em suma, são três formas de executar a busca: 1) pela interface Web do Solr; 2) por execuções de URLs contendo a expressão de busca; e 3) por uma aplicação que consome dados via API.

A interface Web do Apache Solr apresenta a característica de permitir a execução de diversas funções, em que é possível realizar buscas nos documentos anteriormente indexados por meio de uma interface gráfica própria. A Figura 29 exhibe a interface gráfica Web para realizar buscas na ferramenta Apache Solr.

Figura 29 - Interface de busca Apache Solr



Fonte: Elaborada pelo autor

Há diversas opções para especificar a busca da interface gráfica, demonstrada na Figura 29, sendo possível visualizar a variedade de funções que a ferramenta permite, tanto na indexação quanto na busca.

Uma segunda opção para realizar as buscas é por meio de requisições. A requisição ocorre por meio de uma URL, que contém a expressão de busca construída; um exemplo simples de uma busca pode ser visto a seguir: “<http://localhost:8983/solr/dissertacao/select?q=environmental&wt=xml&indent=true>”. No caso dessa URL, o servidor está funcionando localmente, na porta 8983, em uma base de dados chamada de “dissertacao”; também é possível visualizar a expressão de buscas utilizadas, em que a parte que segue os caracteres “?q” representa tal expressão. A solicitação apresenta como retorno um XML com os documentos encontrados que satisfaçam a solicitação.

Ainda há uma terceira opção, em que é utilizado uma API, que permite realizar todas as interações do servidor do Apache Solr com um programa construído em Java. Tal API chama-se Solrj⁶, sendo mantido pela própria Apache. No contexto desse protótipo foi utilizado esse meio para realizar as buscas e as indexações, pela expansão de funções permitida, a partir do uso dessa API.

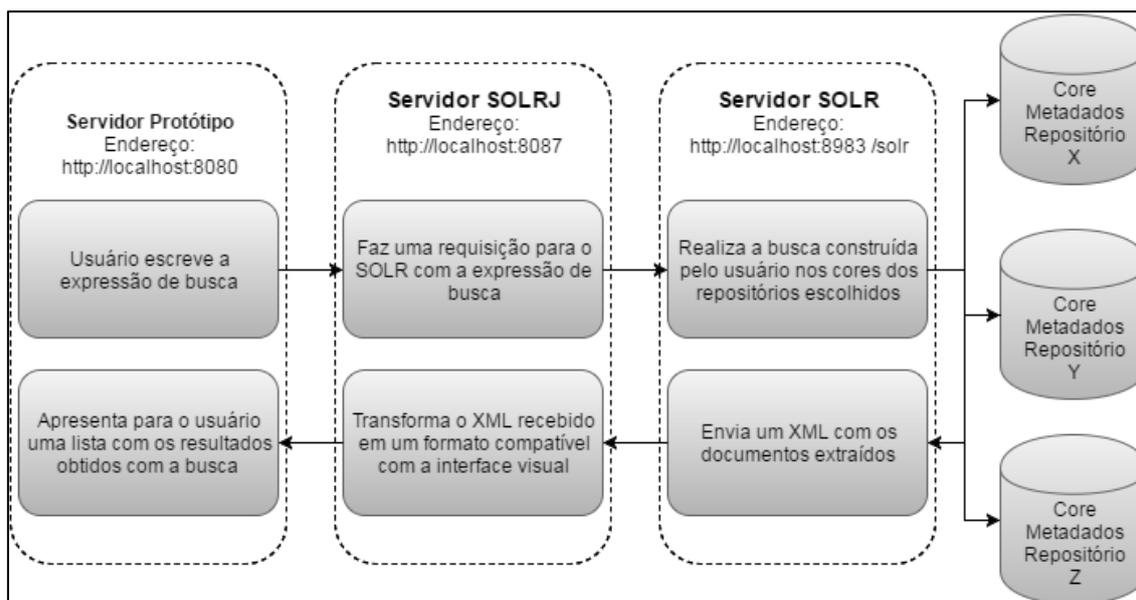
O processo de busca realizado ocorre de modo similar. Assim, o agente controlador envia para o servidor SOLRJ um XML contendo a expressão de busca gerada a partir das interações feitas com o usuário; tal servidor irá transformar a expressão em uma consulta do SOLR, sendo definida, como critério de relevância, uma classificação da própria ferramenta SOLR. O servidor SOLRJ envia para o servidor Apache SOLR a consulta, recebendo um XML, que contém os registros encontrados que satisfaçam a expressão de busca; o servidor SOLRJ envia tal XML para o agente controlador, que irá posteriormente apresentar os registros para o usuário.

Além da expressão de busca construída pelo usuário, o Solrj aponta em quais repositórios será feita a busca do usuário. Assim, o servidor Solr irá recuperar, dentro de cada *core* dos repositórios escolhidos, os registros que atendam à expressão de busca construída pelo usuário.

A Figura 30 apresenta o modo como interagem os elementos citados, adicionando também as URLs utilizadas localmente, para o desenvolvimento desse protótipo. Na mesma figura é possível visualizar como cada passo se relaciona com os outros serviços presentes no protótipo.

⁶ Solrj. Disponível em: <<https://wiki.apache.org/solr/Solrj>>. Acesso em: 07 jan. 2016.

Figura 30 - Funcionamento do módulo de indexação e de busca



Fonte: Elaborada pelo autor

A partir dos processos desenvolvidos, identifica-se que a camada de dados está integrada com as demais camadas do modelo, realizando a execução tanto da busca, por meio da expressão montada com o auxílio das ontologias, quanto da inserção de novas fontes informacionais.

Essa camada é a última camada do modelo, tendo sua importância por realizar a interface com os dados obtidos dos repositórios digitais escolhidos pelos usuários.

6.5 Demonstração de funcionamento

Buscando demonstrar os passos executados pelo protótipo, executou-se um ciclo completo em que todas as partes da implementação são demonstradas sequencialmente. Para isso, utilizou-se um fragmento da ontologia chamada “AKTiveSAOntology”, que apresenta informações de domínio geral.

Tal ontologia foi escolhida por apresentar relações da OWL, possibilitando a demonstração das funcionalidades do sistema. Para o termo de busca inicial, foi definido o termo “*Weapon System*”, devido os repositórios utilizados conterem esse termo em seu *corpus* de documentos e a ontologia apresentar uma classe com esse termo.

Os repositórios digitais escolhidos para a realização da busca foram o Repositório Institucional UNESP e o Repositório Digital LUME da UFRGS (Universidade Federal do Rio Grande do Sul). Esses repositórios foram escolhidos por serem repositórios com

um alto número de registros, sendo referências de estudos publicados, tratando da temática de repositórios institucionais.

As definições explicitadas foram definidas na tela inicial, como pode ser visualizado na figura 31.

Figura 31 - Definição das configurações da demonstração do protótipo

The screenshot shows a web interface titled "BUSCA EM REPOSITÓRIOS" with a dark header containing "Interoperabilidade" and "Início". Below the title is a search bar with the text "Escreva sua pesquisa" and the input "WeaponSystem". Underneath, there are two sections for ontology configuration: "Ontologia - Arquivo" with a button "Escolher arquivo" and the text "AKTiveSAOntologyTeste.owl", and "Ontologia - URL" with an empty input field. A table below lists selected repositories:

Selecionar	Repositório
<input checked="" type="checkbox"/>	LUME
<input checked="" type="checkbox"/>	UNESP

A large blue button labeled "BUSCAR" is positioned below the table. At the bottom of the interface, there is a row of logos for "unesp", "GP-nti", "FAPESP", "CAPES", "USP", and a CD-ROM icon. The footer text reads "Desenvolvido por: Caio Saraiva Coneglian e José Eduardo Santarem Segundo".

Fonte: Elaborada pelo autor.

A partir dessa definição, bem como do processamento das etapas da semântica e do controle relatadas anteriormente, são apresentadas, na sequência, as relações encontradas entre o termo de busca e a ontologia definida. A tela das relações, conforme relatado, compreende as propriedades semânticas, apresentando cada uma delas uma ação dentro do âmbito da recuperação da informação.

A figura 32 apresenta as relações encontradas entre a ontologia e o termo de busca "Weapon System".

Figura 32 - Relações entre o termo “*Weapon System*” e a ontologia definida

Interoperabilidade Início

Relações Busca: Weapon System

Hierarquia

Termo	Selecionar
Thing	<input type="checkbox"/>
Resource	<input type="checkbox"/>

Diferenciação

Termo	Selecionar
Air To Air Missile Release System	<input checked="" type="checkbox"/>
Air To Ground Missile Release System	<input type="checkbox"/>

Igualdade

Termo	Selecionar
Conventional Weapon System	<input checked="" type="checkbox"/>
Thing	<input type="checkbox"/>
Resource	<input type="checkbox"/>

BUSCAR

Desenvolvido por: Caio Saraiva Coneglian e José Eduardo Santarem Segundo

Fonte: Elaborada pelo autor.

Conforme apresentado na figura 32, foram obtidas relações do tipo hierarquia, diferenciação e igualdade. A partir dessas relações, foram escolhidos os termos “*Air to Air Missile Release System*”, com uma relação do tipo de diferenciação, e o termo “*Conventional Weapon System*”, com uma relação de igualdade.

Com as relações escolhidas, o sistema constrói uma nova expressão de busca, tendo como base as ações das propriedades da OWL na recuperação da informação. A nova expressão construída pelo sistema pode ser visualizada na figura 33, contendo diferentes ações, de acordo com as relações escolhidas, tanto referentes à diferenciação, quanto à igualdade.

Figura 33 - Nova expressão de busca gerada

```
(Weapon System OR Conventional Weapon System) NOT (Air to Air
Missile Release System)
```

Fonte: Elaborada pelo autor.

O sistema utiliza a expressão construída, para realizar a busca nos metadados dos repositórios escolhidos (UNESP e LUME-UFRGS), utilizando o Apache SOLR para

realizar a busca. Esse provedor de serviços consulta os conjuntos de dados de ambos os repositórios, retornando os documentos que atendam à expressão de busca. Os resultados obtidos são apresentados na figura 34, que contém a listagem exibida aos usuários.

Figura 34 - Listagem dos resultados obtidos

Interoperabilidade		Início	Novo Repositório	Fazer Login
<h2>Resultados</h2>				
Registros				
[Desenvolvimento e caracterização de nanocápsulas de poli (L-lactídeo) contendo benzocaína] - LUME				
Autor				Data
[Araújo, Daniele Ribeiro de]				[Thu Dec 31 22:00:00 BRST 2009]
Palavras-Chave				SCORE
[drug release]				0.09000171
[Desenvolvimento e caracterização de nanocápsulas de poli (L-lactídeo) contendo benzocaína] - UNESP				
Autor				Data
[Araújo, Daniele Ribeiro de]				[Thu Dec 31 22:00:00 BRST 2009]
Palavras-Chave				SCORE
[drug release]				0.09000171
[Stochastic simulation of time-series models combined with geostatistics to predict water-table scenarios in a Guarani Aquifer System outcrop area, Brazil] - LUME				
Autor				Data
[Tanikawa, Diego H.]				[Wed Oct 31 22:00:00 BRST 2012]
Palavras-Chave				SCORE
[Brazil]				0.08314002

Fonte: Elaborada pelo autor.

Por fim o usuário, ao escolher um registro dentre os apresentados, acessa ao registro completo, com todos os metadados coletados do repositório, permitindo o acesso ao registro e a seu respectivo PDF, na fonte de origem; no caso, ou o repositório UNESP ou o repositório LUME. A tela com o registro completo é visualizada na figura 35.

Figura 35 - Registro completo de um objeto do repositório

Interoperabilidade	Início	Novo Repositório	Fazer Login
--------------------	--------	------------------	-------------

[Desenvolvimento e caracterização de nanocápsulas de poli (L-lactídeo) contendo benzocaína]

Autor:	[Araújo, Daniele Ribeiro de]
Auxílio:	[Universidade Estadual Paulista (UNESP)]
Resumo:	[In this paper we describe the preparation poly (L-lactide) (PLA) nanocapsules as a drug delivery system for the local anesthetic benzocaine. The characterization and in vitro release properties of the system were investigated. The characterization results showed a polydispersity index of 0.14, an average diameter of 190.1 ± 3 nm, zeta potential of -38.5 mV and an entrapment efficiency of 73%. The release profile of Benzocaine loaded in PLA nanocapsules showed a significant different behavior than that of the pure anesthetic in solution. This study is important to characterize a drug release system using benzocaine for application in pain treatment.]
Data:	[Thu Dec 31 22:00:00 BRST 2009]
Palavras-Chaves:	[drug release]
Formato:	[65-69]
Cobertura:	
Identificador:	[S0100-40422010000100013.pdf]
Língua:	[por]
Publicador:	[Sociedade Brasileira de Química]
Relações:	[Química Nova]

Fonte: Elaborada pelo autor.

A partir dessa demonstração, identifica-se que o protótipo possibilita a execução de todos os passos propostos pelo modelo descrito na seção 5, sendo que a seção 6 indica as fases realizadas para a sua implementação.

7 TRABALHOS FUTUROS

A partir dos resultados alcançados na presente dissertação, e das possibilidades que tal pesquisa permitiu vislumbrar, no sentido de tornar as buscas na web cada vez mais eficazes, com maior grau de relevância, desdobramentos delas apresentam-se como trabalhos futuros, para pesquisa e implementação, tais como a aplicação das ações das propriedades da OWL em outros contextos, sendo um possível cenário, a utilização desse modelo em motores de buscas tradicionais, abrindo para elas múltiplas alternativas.

Nesse sentido, pesquisas bastante restritas, pela própria limitação desses motores, onde uma integração maior entre operadores do próprio sistema, impediam um diálogo mais efetivo entre os campos da Recuperação da Informação e da Web Semântica, restringindo as possibilidades de resultados mais satisfatórios, poderiam ser ampliadas e terem realmente uma relevância expressiva.

Outro ponto a ser explorado, referente ao uso das tecnologias da Web Semântica, refere-se ao aprimoramento do motor de SPARQL, realizando a definição prévia de algumas ontologias já construídas como padrão, retirando a responsabilidade da inserção de uma ontologia no sistema ou, dito em outros termos, não havendo mais a necessidade de que ontologias diversas, a cada momento, tivessem que ser inseridas no sistema, visto que elas, ou pelo menos grande parte delas, já figurariam nele como padrão.

No que se refere a interoperabilidade em repositórios digitais, um trabalho futuro possível de realização trata da implementação em casos reais da proposta que permite ao usuário cadastrar e escolher os repositórios utilizados durante a busca. Ainda nesse cenário, é necessário que sejam estabelecidos meios que considerem os diferentes usos que os repositórios fazem do padrão *Dublin Core*, em que ocorre de um mesmo tipo de informação estar descrito em diferentes campos, conforme a aplicação do padrão em cada repositório. O tratamento dessa questão evitaria falhas em apresentar ou em buscar as informações nos campos incorretos, devido a utilização diferente do padrão *Dublin Core*.

Um outro ponto que pode ser expandido trata-se do uso do Apache Solr, aprimorando a utilização dessa ferramenta em domínios da Web Semântica. Nesse sentido, uma maior integração entre as ferramentas de busca e de indexação computacional, caso integradas ao SPARQL e às propriedades da OWL, possibilitaria tornar a busca mais eficiente, quando combinada com as técnicas das ações da OWL na recuperação da informação.

Por fim, porém não menos importante, a utilização de um motor de inferências auxiliando a descoberta do motor de SPARQL poderia levar ao aprimoramento das relações encontradas entre o termo de busca e a ontologia, tornando o processo da criação da expressão de busca mais completo, com um número maior de argumentos obtidos ao explorar a ontologia.

8 CONSIDERAÇÕES FINAIS

A natureza interdisciplinar da Ciência da Informação permite que diversos estudos abarcados por essa área do conhecimento perpassem por outras disciplinas na busca de solucionar questões relativas à informação. Nesse âmbito, a recuperação da informação, estando em seu princípio vinculada à Ciência da Informação, apresenta a necessidade de aprofundar suas pesquisas para que seja capaz de fornecer uma luz no que tange aos usuários encontrarem informações relevantes dentro da Web, ambiente digital com uma quantidade incalculável de informações.

Estudos relativos à relevância dentro da recuperação da informação fornecem um meio para aprimorar a forma como motores de buscas e outros sistemas de recuperação da informação encontrem dados que irão atender às necessidades informacionais dos usuários. Tais pesquisas auxiliam a traçar hipóteses que tentam compreender o comportamento de busca dos usuários, ao passo que visam localizar, dentre um *corpus* de documentos, aqueles que melhor se vinculam ao contexto que um usuário busca.

Não obstante, os estudos de relevância devem percorrer por outros campos do conhecimento, para serem capazes de compreender com um número maior de argumentos as necessidades dos usuários, bem como possuir uma maior contextualização do usuário e dos dados utilizados como fonte de informação.

Um caminho evidente para a resolução de tais questões é a aproximação da Web Semântica com a recuperação da informação. A evolução dos estudos tratando da Web Semântica está criando um arcabouço teórico acerca dos conceitos e das tecnologias dessa proposta, que tem como consequência a execução de estudos teóricos e pragmáticos capazes de trazer contribuições fundamentais em diversos campos de estudos, principalmente no que tange à compreensão do contexto de um determinado domínio.

Essa intersecção entre a recuperação da informação e a Web Semântica tem como objetivo dar mais propriedade para os mecanismos computacionais compreenderem aspectos relativos a uma determinada busca, que por vezes são compreendidos somente por uma pessoa, não sendo traduzido corretamente para as máquinas. Alguns conceitos e tecnologias da Web Semântica, como as ontologias OWL, o modelo de dados RDF e o protocolo de consultas SPARQL podem ter um papel central nessa tarefa, colaborando em aprimorar a recuperação da informação, inserindo uma maior relevância nos resultados obtidos.

Um outro aspecto tratado neste trabalho diz respeito ao processo de recuperação da informação em repositórios digitais. Os repositórios digitais são mecanismos essenciais no contexto atual acadêmico, pela necessidade de preservação e de divulgação da produção intelectual das instituições de ensino, principalmente. Porém o número de trabalhos científicos produzidos está aumentando constantemente, necessitando haver reflexões acerca de meios que permitam aos usuários encontrarem os documentos que atendam às necessidades de informação dos usuários.

Algumas iniciativas, como a promoção de interoperabilidade entre repositórios, criando provedores de serviços, vêm sendo amplamente utilizadas, sendo uma forma de reunir em um único ambiente registros para acesso dos usuários. Essas iniciativas, ainda assim, apresentam problemas, uma vez que o número de registros pode ser extenso, sem que o usuário possa escolher em quais repositórios serão realizadas as buscas.

Diante desse cenário, o presente trabalho buscou apontar um caminho que trace essas correlações entre esses campos de estudos, criando uma forma de utilizar ontologias como uma forma de compreender o contexto em que o usuário cria sua expressão de busca. Para alcançar o objetivo de propor um modelo de recuperação da informação utilizando ontologias, a pesquisa explorou diversos conceitos e tecnologias da Web Semântica, utilizando o cenário de interoperabilidade em repositórios digitais para demonstrar a viabilidade desta proposta.

As relações entre os diversos objetos de estudos utilizados para serem alcançados os objetivos forneceram uma série de resultados alcançados, que perpassam desde a concepção de uma proposta de interoperabilidade de repositórios digitais interativa pela criação de um motor de SPARQL responsável por explorar as ontologias buscando relações com determinados termos, pela definição de ações de cada propriedade da OWL no âmbito da recuperação da informação, chegando até a concepção final do modelo, objetivo principal desta pesquisa.

A definição de uma proposta de interoperabilidade de repositórios digitais interativa foi uma das primeiras definições realizadas neste trabalho, pois foi a partir dessa concepção que foi possível relacionar o modelo que utiliza os conceitos e das tecnologias da Web Semântica com o cenário da pesquisa. Essa fase da pesquisa é a parte importante da camada de dados desta pesquisa, em que há a interação do modelo com os dados dos repositórios digitais.

A proposta de interoperabilidade de repositórios digitais utiliza padrões existentes como o protocolo OAI-PMH, seguindo as bases de provedores de serviços e de

provedores de dados. O destaque dessa fase da pesquisa é a inserção de interatividade na escolha das fontes informacionais (repositórios digitais), no momento em que o usuário realiza a busca. A proposta foi construída visando permitir que o usuário pudesse cadastrar uma nova fonte informacional, de modo que o sistema realizasse a coleta dos metadados do repositório escolhido pelo usuário, permitindo que o usuário, ao realizar uma busca, escolha somente os repositórios em que ele deseja obter informações.

Tal interatividade visa tornar um provedor de serviços mais dinâmico, de forma que o usuário não necessite realizar buscas em diversos repositórios, ou em diversos ambientes de provedores de serviços, que reúnam somente parcialmente o conjunto de registros que ele deseja. Essa proposta indica um novo meio de relacionar os repositórios dentro de um ambiente interoperável, entregando ao usuário uma plataforma em que o controle das fontes utilizadas na recuperação está sob sua responsabilidade, não limitando a busca realizada.

Vale destacar ainda que a implementação desta proposta demonstrou a viabilidade dessa forma de interação entre o sistema e os usuários, bem como mostrou ser factível um sistema que permite o cadastramento dinâmico de novas fontes informacionais (repositórios digitais). Outra característica que foi validada pela implementação foi a escolha, pelos usuários, dos repositórios que seriam utilizados para a realização da busca, no momento da sua execução.

Um segundo ponto concebido nesta proposta foi o motor de SPARQL, parte da camada semântica. Esse motor é responsável pela interação do sistema com a ontologia, realizando os processos referentes à localização das relações existentes entre os termos de busca do usuário e as ontologias de domínio. Em suma, o motor de SPARQL irá localizar o termo de busca do usuário dentro da ontologia, identificando quais são os relacionamentos que esse termo possui com outras classes e propriedades.

Os desafios para a construção desse motor ficaram por conta da execução em tempo real das consultas, que processam os termos de buscas escolhidos, bem como a ontologia de domínio definida pelo usuário. Tal fato fez com que o motor de SPARQL executasse suas tarefas em cada requisição, localizando todas as relações da OWL que estavam vinculadas à busca do usuário, obtendo também as definições de propriedades da OWL vinculadas à cada relação encontrada.

A localização dessas relações permite que sejam extraídos argumentos acerca da busca do usuário, explicitando informações sobre o contexto em que esse termo se insere.

As relações localizadas possuem um nível de semântica formal devido à utilização de ontologias OWL, que possuem propriedades que caracterizam as relações existentes.

Ainda na camada semântica, um dos principais resultados desta pesquisa foi desenvolvido ao realizar-se a definição das ações das propriedades da OWL na recuperação da informação. Essa definição tem como princípio o fato de que as propriedades da OWL revelam características acerca da semântica formal de cada relação, indicando, para os mecanismos computacionais, informações acerca do contexto de cada relação.

Desta feita, essas propriedades da OWL podem ser utilizadas na recuperação da informação, uma vez que podem auxiliar na definição de uma nova expressão de busca, utilizando termos relacionados, com uma expressividade mais elevada. Além do mais, com a obtenção de características da relação de um termo com outro é possível que essa nova expressão de busca relacione tais termos, considerando o significado que aquela determinada relação possui.

Um outro benefício alcançado pela obtenção das propriedades da OWL relacionadas ao termo de busca é a capacidade de explorar a ontologia, utilizando a semântica envolvida em cada relação. Assim, ao localizar as relações que um determinado termo possui, o motor de SPARQL pode utilizar o significado das relações das chamadas propriedades de propriedades, para obter outras relações, localizando novos termos relacionados. A utilização das propriedades da OWL, em conjunto com o protocolo SPARQL, possibilita que a exploração e a extração das relações sejam bem-sucedidas, uma vez que considera o significado de cada relação, obtendo maiores quantidades de relações, com os seus respectivos significados.

A definição das propriedades da OWL dentro da recuperação da informação é fundamental para pesquisas subsequentes, pois possibilitará a inserção de semântica formal no processo de recuperação da informação. Essa fase da pesquisa contempla a principal intersecção entre a Web Semântica e a recuperação da informação, pois permite uma ação clara e objetiva de como as ontologias podem ser utilizadas para aprimorar o processo de recuperação da informação, não sendo utilizadas somente como estruturas de tesouros e de taxonomias. O presente trabalho utiliza as principais funções das ontologias, que correspondem a uma lógica para contextualizar as informações de um domínio.

A camada de nível superior à camada semântica é a camada de controle, em que ocorre a montagem das expressões de busca, de acordo com as relações obtidas na camada semântica. O fato de a montagem da expressão de busca ocorrer em consonância com as

relações encontradas na camada semântica possibilita que o usuário escolha somente aquelas relações que estão de acordo com as suas necessidades de informação. Essa interação entre o usuário e as relações permite que o controle dos termos utilizados fique a cargo do usuário, não criando expressões que estejam distantes da sua necessidade, por uma diferença entre o contexto da ontologia com o contexto do usuário, em algum caso específico.

Vale destacar que a montagem da expressão considera as ações definidas pelas propriedades da OWL na recuperação da informação, pois o tratamento para cada relação encontrada varia de acordo com o significado semântico de cada expressão. Dessa forma, o usuário somente seleciona se aquela relação estiver de acordo com o contexto de sua busca, ficando a cargo do sistema montar a expressão de busca em si, seguindo as ações das propriedades da OWL dentro da recuperação da informação.

A definição desta proposta de montagem de expressão e de sua respectiva implementação demonstra que a interatividade entre o usuário e o sistema aprimora os resultados, uma vez que o sistema não terá a função de descobrir quais daquelas relações encontradas satisfazem as necessidades dos usuários. A função do sistema fica somente em localizar as relações com os seus significados, seguindo uma representação coerente e formal do domínio, no caso as ontologias. Esse esquema retira possíveis ambiguidades do sistema, deixando a cargo do usuário essa definição, enquanto o sistema trata das questões objetivas e lógicas, que são as ontologias, a localização das relações e a montagem da expressão de busca

Por fim, o último passo realizado para atingir o objetivo desta proposta foi a construção da camada de apresentação, que tem como foco ser a camada de interação entre o modelo e o usuário. Essa camada é responsável por apresentar aos usuários os resultados da busca, por possibilitar que o usuário cadastre um novo repositório e por permitir que o usuário selecione as relações semânticas que atendam a sua necessidade de busca.

A camada de apresentação auxilia em demonstrar como os resultados da pesquisa foram atingidos, uma vez que é possível identificar a forma como ocorre a interação entre os usuários e o sistema. Outra questão importante a ser destacada foi o fato de que a implementação possibilitou demonstrar a forma como as relações semânticas podem ser apresentadas aos usuários, no caso, por meio de um agrupamento das relações, de acordo com as definições de divisões de cada propriedade.

Pelo que foi apresentado no texto das conclusões e dos resultados obtidos com as diversas camadas da pesquisa constata-se que o objetivo de propor modelo de recuperação da informação em repositórios digitais, que utilize os conceitos e as tecnologias da Web Semântica para aprimorar a contextualização das necessidades informacionais dos usuários foi alcançado. Tal afirmação é feita, tendo como base a criação do modelo teórico que apresenta as relações entre todas as camadas do modelo, tendo o usuário como agente central da interação com o sistema.

Vale destacar, ainda, que a implementação foi construída como prova de conceito, demonstrando que a proposta criada é passível de ser executada, validando essa forma de inserir a Web Semântica dentro do contexto da recuperação da informação.

Além disso, com os resultados obtidos foi possível identificar que o presente trabalho conseguiu ultrapassar a barreira do uso de ontologias como simples elementos de relações hierárquicas, sem verificar as relações semânticas. Dessa forma, utilizaram-se ontologias analisando as propriedades axiomáticas, que apresentam um nível de semântica formal acima de taxonomias e tesouros, possibilitando uma contextualização do domínio em que o usuário se encontra.

Ademais, a implementação da ferramenta de interoperabilidade demonstra eficiência ao indexar computacionalmente e permitir que metadados de distintos repositórios sejam armazenados em um único ambiente. Tal ferramenta permite expandir a abrangência do módulo semântico, ao trabalhar com um maior número de documentos indexados, tornando, assim, possível a localização dos documentos com maior precisão.

O arcabouço teórico fornecido pela Ciência da Informação, em especial no que tratam as pesquisas de Informação e Tecnologia, foi capaz de promover um diálogo com outras áreas do conhecimento, ao passo que buscou aprofundar o entendimento no tratamento dos dados e das informações no âmbito da Web Semântica. Nesse contexto, a compreensão acerca de ontologias, oriunda da Ciência da Informação, permitiu que as análises realizadas, das propriedades da OWL, estivessem em clara harmonia com os princípios da recuperação da informação.

A união entre as propriedades da OWL e da recuperação da informação possibilita que a relevância da recuperação da informação aumente, pois permite que o contexto do usuário seja trabalhado, numa configuração em que as respostas das buscas considerem outros aspectos, que um sistema sintático é incapaz de fornecer.

Portanto, esta pesquisa avançou significativamente no que corresponde ao uso de ontologias e outras tecnologias e conceitos de Web Semântica, utilizando uma definição concreta de ações que as propriedades da OWL têm dentro da recuperação da informação.

REFERÊNCIAS

- 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.
- ALVES, R. C. V.; SANTOS, P. L. V. A. da C. Metadados. In: _____. **Metadados no domínio bibliográfico**. Niterói: Intertexto, 2013. p.35-64.
- ANTONIOU, G.; VAN HARMELEN, F. **A Semantic Web Primer**. MIT press, 2004.
- APACHE. **Solr**. 2014. Disponível em: <<http://lucene.apache.org/solr/>>. Acesso em: 12 jan. 2016.
- APACHE. **Apache Lucene Scoring**. 2012. Disponível em: <https://lucene.apache.org/core/3_6_0/scoring.html>. Acesso em: 17 out. 2016.
- APACHE, **Lucene**. 2011-12. Disponível em: <<https://lucene.apache.org/core/>>. Acesso em: 12 jan. 2016.
- ASSUMPCÃO, F. S.; SANTOS, P. L. V. A. da C. Representação no domínio bibliográfico: um olhar sobre os Formatos MARC 21. **Perspectivas em Ciência da Informação**, v. 20, n. 1, p. 54-74, 2015. Disponível em: <<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2054>>. Acesso em: 2 ago. 2016.
- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação: conceitos e tecnologia das máquinas de busca**, 2.ed. Porto Alegre: Bookman, 2013.
- BECHHOFFER, S. et al. **OWL Web Ontology Language reference**. 2004. Disponível em: <<http://www.w3.org/TR/owl-ref/>>. Acesso em: 19 dez. 2015.
- BERNERS-LEE, T. **Information management: A proposal**. 1989.
- BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic Web. **Scientific American**, v. 284, n. 5, p. 28-37, 2001.
- BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. **Uniform resource identifier (URI): Generic syntax**. 2004.
- BRAY, T. et al. Extensible markup language (XML) 1.1 (Second Edition). **World Wide Web Consortium Recommendation**. 2006. Disponível em: <<http://www.w3.org/W3cSpec/XML/2/REC-xml11-20060816.pdf>>. Acesso em: 9 dez. 2015.
- BREITMAN, K. K. **Web semântica: a internet do futuro**. Rio de Janeiro: LTC, 2005. 190 p.
- BRESLIN, J. G.; PASSANT, A.; DECKER, S. **The Social Semantic Web**. Springer Science & Business Media, 2009.

BORKO, H. Information science: what is it? **American documentation** 19.1, p. 3-5, 1968. Disponível em:
<<http://onlinelibrary.wiley.com/doi/10.1002/asi.5090190103/abstract>>. Acesso em: 29 fev. 2016.

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 Twente, Enschede, 1997. Disponível em:
<<http://doc.utwente.nl/17864/1/t0000004.pdf>>. Acesso em: 10 jan. 2016.

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 NUUVENS, 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: 2 fev. 2016.

CARDOSO, O. N. P. Recuperação de Informação. **INFOCOMP Journal of Computer Science**, v. 2, n. 1, p. 33-38, 2004. Disponível em:
<<http://www.dcc.ufla.br/infocomp/index.php/INFOCOMP/article/view/46>>. Acesso em: 15 fev. 2016.

CASTRO, F. F.; SANTOS, P. L. V. A. C. Elementos de interoperabilidade na perspectiva da catalogação descritiva. **Informação & Sociedade**. João Pessoa, v. 24, n. 3, p. 13-25, set/dez. 2014. Disponível em:
<<http://www.ies.ufpb.br/ojs/index.php/ies/article/view/16660>> Acesso em: 02 dez. 2015.

CATARINO, M. E.; BAPTISTA, A. A. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramaZero-Revista de Ciência da Informação**, v. 8, n. 3, 2007. Disponível: http://www.dgz.org.br/jun07/Art_04.htm

CATARINO, M. E.; SOUZA, T. B. A representação descritiva no contexto da Web Semântica. **Transinformação**, Campinas, v. 24, n. 2, p. 77-90, 2012.
DAS NEVES, Teodora Marly Gama. Livre acesso à publicação acadêmica. **Ciência da Informação**, v. 33, n. 3, p. 116-121, 2004.

DECKER, S. et al. The semantic Web: The roles of XML and RDF. **Internet Computing, IEEE**, v. 4, n. 5, p. 63-73, 2000.

DIAS, T. D.; SANTOS, N. Web Semântica: Conceitos Básicos e Tecnologias Associadas. **Cadernos do IME-Série Informática**, v. 14, p. 80-92, 2013.

DUCHARME, B. **Learning Sparql**. O'Reilly Media, Inc., 2011.

DZIEKANIAK, G. V.; KIRINUS, J. B. Web semântica. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 9, n. 18, p. 20-39, 2004.

DCMI. **Dublin Core Qualifiers**. 2000. Disponível em: <<http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/>> Acesso em: 13 jan. 2015.

DCMI. **Dublin Core Metadata Element Set**, Version 1.1. 2012. Disponível em: <<http://dublincore.org/documents/dces/>>. Acesso em: 2 dez. 2015.

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.

FERREIRA, J. A. **Wikis semânticos: da Web para a Web Semântica**. 2014. 131 f. Dissertação (Mestrado em Ciência da Informação). Faculdade de Filosofia e Ciências – Universidade Estadual Paulista, Marília, 2014. Disponível em: <<http://hdl.handle.net/11449/108380>>. Acesso em 8 fev. 2016.

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.ies.ufpb.br/ojs/index.php/ies/article/view/15436/9681>> Acesso em: 15 out. 2015.

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.

GARCIA, P. D. A. B.; SUNYE, M. S. O. protocolo OAI-PMH para interoperabilidade em bibliotecas digitais. In: **CONGRESSO DE TECNOLOGIAS PARA GESTÃO DE DADOS E METADADOS DO CONE SUL**. 2003.

GRÁCIO, J. C. A. **Metadados para a descrição de recursos da Internet: o padrão Dublin Core, aplicações e a questão da interoperabilidade**. 2002. Tese (Doutorado em Ciência da Informação) -Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2002.

GRIMALDO, W. A. M. Sociedad de la información: metadatos y futuro de la Internet en la recuperación de información de calidad. **Bibliotecas & Tecnologías de La Información**. 2004. Disponível em: <http://eprints.rclis.org/7010/1/Art%C3%ADculo_Sociedad_de_la_Informaci%C3%B3n_y_Metadatos.pdf>. Acesso em: 14 dez. 2015.

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: 12 dez. 2015.

GUARINO, N. Formal ontology in information systems. In: **Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy**. IOS press, 1998.

HITZLER, P.; KROTZSCH, M.; RUDOLPH, S. **Foundations of semantic Web technologies**. CRC Press, 2009.

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/wo11/doi/10.1002/asi.21261/full>>. Acesso em: 10 nov. 2016.

IBICT. **Sobre Repositórios Digitais**. Instituto Brasileiro de Informação em Ciência e Tecnologia. 2015. Disponível em: <<http://www.ibict.br/informacao-para-ciencia-tecnologia-e-inovacao%20/repositorios-digitais>> Acesso em 10 fev. 2016.

JACOB, E. K. Ontologies and the semantic Web. **Bulletin for the American Society for Information Science and Technology**, v. 29, n.4, p.19-22, Abr/Mai 2003. Disponível em: <<http://asis.org/Bulletin/Apr-03/Jacob.pdf>>. Acesso em: 12 dez. 2015.

JAIN, V.; SINGH, M. Ontology Based Information Retrieval in Semantic Web: A Survey. **International Journal of Information Technology and Computer Science**, v. 5, n. 10, p. 62–69, 1 set. 2013.

JANAITE NETO, J.; FERNEDA, E. Ontologia como recurso de padronização terminológica no processo de recuperação de informação. **Informação em Pauta**, Fortaleza, v. 1, n. 1, p. 30-45, jan./jun. 2016. Disponível em: <<http://www.periodicos.ufc.br/index.php/informacaoempauta/article/view/2967/2692>>. Acesso em: 15 nov. 2016.

JONES, C. B.; ALANI, H.; TUDHOPE, D. **Geographical Information Retrieval with Ontologies of Place**. In: MONTELLO, D. R. (Ed.). *Spatial Information Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. 2205p. 322–335.

LAGOZE, C.; VAN DE SOMPEL, H. The Open Archives Initiative: Building a low-barrier interoperability framework. In: **Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries**. ACM, 2001. p. 54-62.. Disponível em <http://www.elbib.ru/index/index.phtml?page=elbib/eng/journal/2001/part6/LS>. Acesso em 19 ago. 2015.

LASSILA, O.; MCGUINNESS, 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: 18 dez. 2015.

LEITE, F.C.L. et al. **Como gerenciar e ampliar a visibilidade da informação científica brasileira: repositórios institucionais de acesso aberto**. IBICT, Brasília. 2009.

LEITE, F. C. L.; ARELLANO, M. A. M.; MORENO, F. P. Acesso livre a publicações e Repositórios Digitais em ciência da informação no Brasil. **Perspectivas em ciência da informação**, 11.1: 82-94. Belo Horizonte. 2006.

LIBRARY OF CONGRESS. **MARC Standards**. 2015. Disponível em: <<http://www.loc.gov/marc>>. Acesso em: 29 fev. 2016.

MARCONDES, C. H.; SAYÃO, L. F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, v. 30, n. 3, p. 24-33, 2001.

MARCUSCHI, L. A. O hipertexto como um novo espaço de escrita em sala de aula. **Revista Linguagem & Ensino**, v. 4, n. 1, p. 79-111, 2001.

MILLER, E. An Introduction to the Resource Description Framework. **D-Lib Magazine**, v. 4, n.5, May, 1998. Disponível em: <<http://www.dlib.org/dlib/may98/miller/05miller.html>>. Acesso em: 8 fev. 2016.

MILLER, E. **The semantic Web**. 2004. Disponível em: <<http://www.w3.org/2004/Talks/0120-semWeb-umich/Overview.html>>. Acesso em: 8 dez. 2015.

MODESTO, L. R. **Representação e Persistência para acesso a Recursos Informacionais Digitais gerados dinamicamente em sítios oficiais do Governo Federal**. 2013. 103 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2013.

PANSANI JUNIOR, E. A. **Ontologias no processo de indexação automática de documentos textuais**. 2016. 126 f. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2016. Disponível em: <<http://hdl.handle.net/11449/138961>>. Acesso em: 2 nov. 2016.

PETINARI, V. S. Repositórios digitais e sua colaboração para Disseminação da produção científica da graduação. In: Seminário Nacional de Bibliotecas Universitárias, São Paulo, 2008. **Anais...** São Paulo: CRUESP, 2008. Disponível em: <<http://www.sbu.unicamp.br/snbu2008/anais/site/pdfs/2878.pdf>>. Acesso em 1 set.2014.

PICKLER, M. E. V. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em Ciência da Informação**, v. 12, n. 1, p. 65-83, 2007.

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 8 fev. 2016.

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.

ROSETTO, M. **Metadados e formatos de metadados em sistemas de informação: caracterização e definição**. 2003. Dissertação (Mestrado em ciências da comunicação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo.

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.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a 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. 2016.

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.

SANTOS, P. L. V. A. C.; ALVES, R. C. V. Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0. **DataGramZero**, Rio de Janeiro, v. 10, n. 6, dez. 2009. Disponível em: <http://www.dgz.org.br/dez09/Art_04.htm>. Acesso em: 8 dez. 2015.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da informação**, v. 24, n. 1, p. 36-41, 1995. Disponível em: <<http://www.brapci.ufpr.br/documento.php?dd0=0000005946&dd1=59269>>. Acesso em: 10 dez. 2015.

SAYÃO, L. F. et al. Software livres para repositórios institucionais: alguns subsídios para a seleção. In: **Implantação e gestão de repositórios institucionais**. EDUFBA, Salvador. 2009.

SHINTAKU, M. **Federação de Repositórios Científicos: identificação, análise e proposta de modelo baseado nas tendências tecnológicas e da Ciência**. 2015. 268 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília. 2015.

SILVA, G. C.; LIMA, T. S. RDF e RDFS na infra-estrutura de suporte a Web Semântica. **Revista Eletrônica de Iniciação Científica**, Porto Alegre, v.2, n.2, mar. 2002. Sociedade Brasileira de Computação. Disponível em: <<http://www2.ic.uff.br/~gsilva/slreic.pdf>>. Acesso em: 11 dez. 2015.

SHAH, U.; FININ, T.; JOSHI, A. Information retrieval on the semantic web. In: **Anais...** ACM Press, 2002. Disponível em: <<http://portal.acm.org/citation.cfm?doid=584792.584868>>. Acesso em: 13 set. 2016.

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: 12 dez. 2015.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, 2004.

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, 2006. Disponível em: <<http://www.scielo.br/pdf/%0D/pci/v11n2/v11n2a02.pdf>>. Acesso em: 15 dez. 2015.

VIANA, C. L. M.; ARELLANO, M. A. M; SHINTAKU, M. **Repositórios institucionais em ciência e tecnologia**: uma experiência de customização do DSpace. 2013.

W3C. **Resource Description Framework (RDF) Model and Syntax Specification**. 1999. Disponível em: <<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>>. Acesso em: 9 dez. 2015.

W3C. **URIs, URLs, and URNs**: Clarifications and Recommendations 1.0. 2001. Disponível em: <<http://www.w3.org/TR/uri-clarification/>> Acesso em: 22 ago. 2015.

W3C. **Resource Description Framework (RDF) Concepts and Abstract Syntax**. 2004. Disponível em: <<http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Structute>>. Acesso em: 9 dez. 2015.

W3C. **Layer Cake**. 2007. Disponível em: <<https://www.w3.org/2007/03/layerCake.png>>. Acesso em: 29 fev. 2016.

W3C. **SPARQL Query Language for RDF**. 2008. Disponível em: <<http://www.w3.org/TR/rdf-sparql-query/>>. Acesso em: 21 dez. 2015.

W3C. **W3C Opens Data on the Web with SPARQL**. 2008b. Disponível em: <<http://www.w3.org/2007/12/sparql-release.html.en>>. Acesso em: 21 dez. 2015.

W3C. **PICS The Platform for Content Selection Home Page**, 2009. Disponível em: <<http://www.w3.org/PICS>>. Acesso em:

W3C. **Web Semântica**. 2011. Disponível em: <<http://www.w3c.br/Padroes/WebSemantica>> Acesso em: 7 dez. 2015.

W3C. **OWL Web Ontology Language**. 2012. Disponível em: <<http://www.w3.org/2001/sw/wiki/OWL>> Acesso em: 19 dez. 2015.

W3C. **OWL 2 Web Ontology Language**. Quick Reference Guide (Second Edition). 2012b. Disponível em: <<http://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/>>. Acesso em: 19 dez. 2015.

W3C. **SPARQL 1.1 Query Language**. 2013. Disponível em: <<http://www.w3.org/TR/sparql11-query/>> Acesso em: 22 ago. 2015.

W3C. **RDF**. 2014. Disponível em: <<http://www.w3.org/RDF>>. Acesso em: 9 dez. 2015.

W3C. **RDF Schema 1.1**. 2014b. Disponível em: <<http://www.w3.org/TR/rdf-schema>>. Acesso em: 11 dez. 2015.

W3C. **Extensible Markup Language (XML)**. 2015. Disponível em: <<http://www.w3.org/XML/>> Acesso em: 9 dez. 2015.

ZENG, M. L.; QIN, J. **Metadata**. Neal-Schuman Publishers, Inc., 2008.