Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Marcelo Gomes Rodrigues

OntoFeed: um Leitor de *Feeds* com Extensão Ontológica

Rio de Janeiro

2011

Marcelo Gomes Rodrigues

OntoFeed: um Leitor de *Feeds* com Extensão Ontológica

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Engenharia de Computação, da Universidade do Estado do Rio de Janeiro. Área de concentração: Geomática.

Orientadora: Prof.ª Dr.ª Neide dos Santos

Coorientador: Prof. Dr. João Araújo Ribeiro

Rio de Janeiro

2011

CATALOGAÇÃO NA FONTE

UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

R696 Rodrigues, Marcelo Gomes.
OntoFeed – um leitor de Feeds com extensão ontológica. / Marcelo Gomes Rodrigues. - 2011.
103f.

Orientadora: Neide dos Santos.
Coorientador: João Araújo Ribeiro
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Web semântica – Teses. 2. Internet – Teses. 3. Engenharia de Computação. I. Santos, Neide dos. II. Universidade do Estado do Rio de Janeiro. III. Título.

CDU 004.738.5:001.102

_____ _____

Assinatura Data

Marcelo Gomes Rodrigues

OntoFeed: um leitor de feeds com extensão ontológica

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Engenharia de Computação, da Universidade do Estado do Rio de Janeiro. Área de concentração: Geomática.

Aprovado em: 23 de agosto de 2011.

Banca Examinadora:

_____

Prof.ª Dr.ª Neide dos Santos (Orientadora)
Instituto de Matemática e Estatística - UERJ

_____

Prof. Dr. João Araújo Ribeiro (Coorientador)
Faculdade de Engenharia - UERJ

_____

Prof. Dr. Orlando Bernardo Filho
Faculdade de Engenharia - UERJ

_____

Prof.ª Dr.ª Fernanda Cláudia Alves Campos
Universidade Federal de Juiz de Fora - UFJF
Departamento de Ciência da Computação

_____

Prof.ª Dr.ª Adriana Aparicio Sicsu Ayres do Nascimento
Centro Universitário Estadual da Zona Oeste – UEZO
Departamento de Ciência da Computação

Rio de Janeiro

2011

DEDICATÓRIA

Aos meus amados pais

João e Marly

AGRADECIMENTOS

Ao Senhor meu Deus, razão do meu viver, pela coragem e esforço com os quais me abençoou, dando-me graça para chegar até aqui. Obrigado Senhor, por tuas imensas provisões!

Aos meus pais João e Marly, por me fornecerem com suas atitudes, os valores da compreensão, do auxílio e do amor. Gestos que me mostraram os verdadeiros fundamentos da vida e propiciaram-me condições para superar dificuldades, e sempre avançar em direção aos meus objetivos, sem nunca desistir.

Aos meus amados orientadores: professora Neide e professor João, pela paciência, pelo carinho, apoio e interesse nas horas mais difíceis. Seria complicado listar o tamanho desta amizade e dos auxílios com os quais me enriqueceram fornecendo-me suas orientações ao longo destes quase três anos. São meus professores e continuarão sendo, mas também preciosos amigos.

Ao querido professor Orlando, por seu apoio durante todo o curso, o que muito me ajudou a conseguir terminar esta importante etapa de minha vida. Ao Cel. Jorge Lucente e toda equipe da Defesa Civil do Estado do Rio de Janeiro, pelas orientações e explicações sobre o Manual de Desastres e o *modus operandi* da instituição.

Ao dileto amigo Wilter Monteiro por me esclarecer em diversos itens do trabalho. Ao meu caro amigo Cláudio Rapello, pelas colaborações valiosas e por todo esmero com que me auxiliou com sua experiência, sabedoria e seu desejo incólume de ajudar.

Aos meus caros amigos Ailton, Angel, Robson, Newton, José Paulo, Herval, Rodrigo, Marco Borsoni e os Alexandres que participaram das minhas aflições e alegrias durante todo o tempo do mestrado. Aos professores e funcionários do mestrado UERJ, com quem tive a honra de estar ao longo deste período.

Não há semântica que consiga traduzir em palavras o que de fato todos os senhores fizeram. Acomodar-me-ei em escolher apenas duas palavras para isto: muito obrigado!

"De sorte que haja em vós o mesmo sentimento que houve também em Cristo Jesus,Que, sendo em forma de Deus, não teve por usurpação ser igual a Deus, Mas esvaziou-se a si mesmo, tomando a forma de servo, fazendo-se semelhante aos homens; E, achado na forma de homem, humilhou-se a si mesmo, sendo obediente até à morte, e morte de cruz. Por isso, também Deus o exaltou soberanamente, e lhe deu um nome que é sobre todo o nome; Para que ao nome de Jesus se dobre todo o joelho dos que estão nos céus, e na terra, e debaixo da terra, e toda a língua confesse que Jesus Cristo é o Senhor, para glória de Deus Pai."

Paulo, Filipenses 2.5-11

RESUMO

RODRIGUES, Marcelo Gomes. *OntoFeed – um leitor de Feeds com extensão ontológica*. 2011. 103f. Dissertação (Mestrado em Engenharia de Computação) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2011.

O problema que justifica o presente estudo refere-se à falta de semântica nos mecanismos de busca na *Web*. Para este problema, o consórcio W3 vem desenvolvendo tecnologias que visam construir uma *Web* Semântica. Entre estas tecnologias, estão as ontologias de domínio. Neste sentido, o objetivo geral desta dissertação é discutir as possibilidades de se imprimir semântica às buscas nos agregadores de notícia da *Web*. O objetivo específico é apresentar uma aplicação que usa uma classificação semi-automática de notícias, reunindo, para tanto, as tecnologias de busca da área de recuperação de informação com as ontologias de domínio. O sistema proposto é uma aplicação para a *Web* capaz de buscar notícias sobre um domínio específico em portais de informação. Ela utiliza a API do Google Maps V1 para a localização georreferenciada da notícia, sempre que esta informação estiver disponível. Para mostrar a viabilidade da proposta, foi desenvolvido um exemplo apoiado em uma ontologia para o domínio de chuvas e suas consequências. Os resultados obtidos por este novo *Feed* de base ontológica são alocados em um banco de dados e disponibilizados para consulta via *Web*. A expectativa é que o *Feed* proposto seja mais relevante em seus resultados do que um *Feed* comum. Os resultados obtidos com a união de tecnologias patrocinadas pelo consórcio W3 (XML, RSS e ontologia) e ferramentas de busca em página *Web* foram satisfatórios para o propósito pretendido. As ontologias mostram-se como ferramentas de usos múltiplos, e seu valor de análise em buscas na *Web* pode ser ampliado com aplicações computacionais adequadas para cada caso. Como no exemplo apresentado nesta dissertação, à palavra “chuva” agregaram-se outros conceitos, que estavam presentes nos desdobramentos ocasionados por ela. Isto realçou a ligação do evento “chuva” com as consequências que ela provoca - ação que só foi possível executar através de um recorte do conhecimento formal envolvido.

Palavras-Chave: Web semântica. Ontologia. Busca semântica. Agregadores de notícia.

ABSTRACT

The problem addressed in this work refers to the lack of semantics in Web search engine. As solution, the W3 consortium has been developing technologies that aim to build a Semantic Web, including the domain ontology. Considering this issue, the work main goal is to discuss the possibilities of placing semantics – context – in the searches in Web feed applications. The specific goal is to propose a Web application that uses a semi-automatic classification of news, by joining information retrieval technologies and domain ontology. The software is able to get news about a given domain from Web information portals. It uses the Google Map API VI for gather the new geo-referenced location, whenever this information is available. To show the proposal feasibility, an example was developed supported by an ontology in the domain of rainfall and its consequences. The results of this new ontology-based feed are allocated in a database e make available for query via the Web. It is expected that the proposed feed offers more relevant results than the current feeds. In addition, the union of technologies sponsored by the W3C and traditional search methods on Web pages were satisfactory for the intended purposes. Ontology is showed as multi-use tool and its value in Web search can be extended for appropriate computer applications. In the example presented, other concepts were added to the word “rainfall”, which is present in the deployments caused by it. This highlighted the connection of the event rainfall with its consequences, action that was only possible to run through a cutout of the formal knowledge involved.

Keywords: Semantic web. Ontology. Semantic search. Feeds.

LISTA DE FIGURAS

LISTA DE ILUSTRAÇÕES

SUMÁRIO

INTRODUÇÃO

As formas mais costumeiramente usadas para armazenar informações são os textos escritos e os digitais. Atualmente, os textos digitais de um grande número de revistas, enciclopédias, livros e jornais podem ser encontrados mediante o acesso à Internet, ainda que o usuário necessite realizar algum tipo de assinatura para lê-los. Uma vez que não somente empresas, governos e universidades, mas também pessoas comuns tornaram-se editoras de conteúdo *Web*, a Internet foi transformada no mais extenso repositório disponível de dados.

Mesmo utilizando filtros de busca para agilizar sua pesquisa e localizar por palavras-chave a informação desejada, o usuário enfrenta problemas de localização dado o grande volume de informações disponíveis. A dificuldade deve-se ao fato da *Web* possuir um ambiente sem padronização, que se assemelha a um caos de dados, devido, entre outras razões, à liberalidade cultural, à informalidade com que os conteúdos são propagados, ao desprovimento de regras para montagem dos documentos e páginas e à falta de censura. Este enorme caudal informativo sem uma estrutura de normas definida é fator gerador de redundâncias e inconsistências de dados.

Uma simples consulta aos dados alocados na *Web* é realizada, na maioria das vezes, escrevendo-se palavras-chave em algum mecanismo de busca. Então, o sistema retorna ao usuário aqueles documentos que seu algoritmo conseguir localizar como compatíveis com o que foi inicialmente digitado. Todavia, nem sempre o retorno devolvido pelo sistema possui a relevância pretendida pelo usuário.

Um dos motivos para que isto ocorra é que a mesma palavra pode ter significados diferentes dependendo do contexto onde ela estiver escrita. Por exemplo, ao se digitar o termo manga no Google, a pesquisa apresenta, dos dez primeiros itens, nove deles relacionados a “manga” (história em quadrinhos de origem japonesa) e um referente a um site de humor. Não aparecem: manga como fruta, como parte de roupa, o goleiro da seleção brasileira de futebol de 1966 nem o sobrenome Manga.

Uma tentativa de solucionar as ambiguidades e os problemas referentes ao significado dos termos procurados é se implementar a busca semântica. Segundo Berners-Lee, Handler e Lassila (2001), para que ela se torne uma realidade, em meio à heterogeneidade de dados e linguagens existentes na Internet, será necessário incorporar significados aos conteúdos e páginas da *Web*, formando assim a *Web* Semântica. A *Web* Semântica funciona como uma extensão da *Web* atual, tendo como motivação principal a cooperação entre homens e

computadores. No entanto, permitir que as palavras sejam entendidas pelas máquinas e sistemas em seus sentidos originais como quando foram escritas e digitadas pelo ser humano é uma tarefa complexa.

Visando constituir padrões para a vastidão dos domínios *Web* e assim estabelecer uma base para a *Web* Semântica, World Wide *Web* Consortium (W3C) vem trabalhando para desenvolver linguagens e tecnologias que possam ser usadas na construção de um modelo capaz de gerar interação entre máquinas. Em seu repertório tecnológico, destaca-se o uso de agentes e linguagens como XML, RDF, OWL e RSS. XML.

- XML (*eXtensible Markup Language)* facilita compartilhamento de informações via *Web*
- RDF (*Resource Description Framework)* trabalha com representação da informação na Internet
- OWL (*Web Ontology Language)* define e instancia ontologias
- RSS (*Really Simple Syndication)* é um agregador de conteúdos XML

Tanto linguagens quanto agentes são facilitadores para a concretização da *Web* Semântica. Por meio do fornecimento de descrições para os dados, baseadas nos padrões do W3C, os computadores identificarão a representação de um termo (nome) com seu real significado, e com isto permitirão que programas agentes possam executar procedimentos e tarefas sobre os conteúdos da Internet.

O grande desafio é transformar o dado em informação e a partir daí gerar conhecimento. Uma das formas de representar conhecimento são as ontologias. O termo ontologia está associado ao modelo de dados que estabelece a representação de conceitos (classes) dentro do contexto de um domínio e engloba os relacionamentos entre eles. Segundo Gruber (1993), uma ontologia é uma especificação explícita e formal de conceitos compartilhados. Fundamentalmente, o objetivo da ontologia é facilitar a elaboração de um domínio modelado pela representação de termos componentes de um vocabulário e as relações destes entre si. Pela ontologia, o dado estaria revestido de um aspecto semântico, pois se localizaria dentro de um domínio específico do conhecimento. As ontologias, alicerçadas nas linguagens do W3C, constituem o alicerce para os mecanismos de busca semântica.

Uma das formas já em uso na *Web* de se realizar consultas específicas em portais e

blogs pré-definidos pelo usuário são as classificações semi-automáticas de notícias. Seus executores, os *Feeds*, são mecanismos de busca, baseados em RSS, uma linguagem que serve como agregadora de conteúdos XML. Através deles, os usuários recebem as atualizações das páginas nas quais se cadastraram, diretamente em seus computadores, sem ter que buscá-las uma a uma nos *sites* de origem. Mesmo agindo sobre páginas *Web* pré-estabelecidas pelo usuário e possuindo recursos de busca *booleana e por rating*, os *Feeds* não contemplam uma busca com contexto, o que poderia ser obtido com o uso de coleções de ontologias de domínio.

O presente trabalho propõe ampliar a relevância dos resultados obtidos pelos *Feeds*, agregando ao programa os recursos da definição de uma ontologia. Conceitos, propriedades e relacionamentos de um domínio específico servirão de base para as buscas nos portais de notícias. Isto permitirá ao usuário selecionar não somente o local das informações, como também carregar o sistema com os termos da ontologia. A expectativa é que o *Feed* proposto seja mais relevante em seus resultados do que um *Feed* comum.

Justificativa

O grande desafio da *Web* atual é acoplar semântica aos seus conteúdos. Esta dissertação trata este problema utilizando como exemplo os *Feeds*. Os *Feeds* possuem um campo chamado descrição, onde há o resumo da notícia enviada pelo portal. Atualmente, o usuário do serviço limita-se a ler as informações contidas neste item. Com a extensão proposta por este trabalho, pretende-se extrair conhecimento do campo descrição através de uma aplicação computacional com base em uma ontologia de domínio. Para demonstrar a proposta, um exemplo foi desenvolvido no domínio de chuvas e suas consequências, que analisará o texto da descrição do *Feed* e fornecerá ao usuário elementos novos (adicionados à notícia), provenientes do contexto da ontologia. Assim, o usuário do serviço de Feeds não somente lerá a notícia, mas também receberá nela as inferências formadas pelo uso da ontologia.

Objetivos do Trabalho

O objetivo geral desta dissertação é discutir as possibilidades de se imprimir semântica às buscas nos agregadores de notícia da *Web*. O objetivo específico é apresentar uma aplicação que usa uma classificação semi-automática de notícias, reunindo, para tanto,

as tecnologias de busca da área de recuperação de informação com as ontologias de domínio. Para mostrar a viabilidade da proposta, foi desenvolvido um exemplo apoiado em uma ontologia para o domínio de chuvas e suas consequências. Os resultados obtidos por este novo *Feed* de base ontológica são alocados em um banco de dados e disponibilizados para consulta via *Web*.

Para atingir este objetivo, os seguintes passos são necessários:

a) Operar classificação semi-automática de notícias (*Feed*), com a base ontológica previamente definida.
b) Construir um banco de dados, onde as informações obtidas pela busca do *Feed* serão armazenadas.
c) Construir um mecanismo para a localização de notícias.
d) Construir um ambiente na *Web* para permitir consulta ao banco de dados deste serviço.
e) Construir ontologias de domínio, no exemplo desenvolvido, domínio “chuvas e suas consequências”.

Organização do Trabalho

Este documento está organizado em três capítulos, além desta introdução. No capítulo 2, são mostradas as tecnologias e linguagens patrocinadas pelo W3C, que visam tornar a *Web* mais semântica. É descrito, também, o funcionamento dos leitores de Feeds e suas características básicas. A seguir, discutem-se ontologia, seus tipos, vantagens de seu uso e as metodologias para sua construção. Finalmente, os trabalhos relacionados às classificações automáticas de notícias são apresentados. O capítulo 3 apresenta o sistema OntoFeed, baseado na linguagem livre PHP, e suas filtragens, associações e classificações automáticas. O capítulo detalha ainda a construção da ontologia “chuvas e suas consequências”. No capítulo 4 são apresentadas as considerações e conclusões finais e os trabalhos futuros.

1 FUNDAMENTAÇÃO TEÓRICA

O problema que justifica o presente estudo refere-se à falta de contexto – ou semântica, nos mecanismos atuais de busca na *Web*. Este problema nos remete aos esforços empreendidos pelo W3 para viabilizar uma *Web* Semântica. Este capítulo oferece a fundamentação teórica que sustenta a proposta da dissertação. De início, ele oferece uma visão geral da *Web* Semântica e discute brevemente recuperação da informação. Os serviços de *feeds* são então apresentados e discutidos e o conceito de ontologias, no contexto dos *feeds* é introduzido.

1.1 *Web* Semântica

Atualmente, a Internet é a maior rede que interliga computadores ao redor do globo. Um de seus pontos fortes é a vasta quantidade de dados disponibilizados. Paradoxalmente, este também é o seu calcanhar de Aquiles, em virtude da falta de padronização, e localizar exatamente o que se quer no universo da *Web*, nem sempre é tarefa simples. (RAMALHO; VIDOTTI; FUJITA, 2005, p.4).

> [...]apresenta-se como desafio a necessidade de singularização contextual, na reconstrução do conhecimento, determinando requisitos de qualidade e relevância dos conteúdos, tornando-se necessário a utilização de categorias que permitam organizar, de maneira eficiente, o "oceano" de dados disponíveis, permitindo a identificação daquilo que realmente interessa ao usuário num contexto preciso. Na realidade tais categorias já existem, porém estas são compreendidas apenas pelos humanos, e não possuem nenhum sentido lógico para os programas de computador.

Dois fatores contribuem fortemente para que a procura por informações na *Web* seja um serviço trabalhoso: (a) a informação não possui um significado bem definido, que proporcione a colaboração e o entendimento entre pessoas e os agentes de software; e, (b) sua forma de organização de dados e informações baseia-se na perspectiva humana, utilizando linguagem natural e HTML (*HyperTex Markup Language*).

W3C, criado em 1994 reunindo empresas e universidades, visa desenvolver novas tecnologias, que se tornem padrão para a Internet e assim viabilizem uma maior organização de seu conteúdo. Um dos projetos do consórcio é a *Web* Semântica (WS). A linguagem HTML usada para escrever páginas na *Web* é limitada para descrever os dados, servindo para dizer como a página deve ser visualizada pelos navegadores. O projeto da *Web* Semântica visa

construir categorias e linguagens que permitam aos computadores obter o sentido dos dados. Ela é uma extensão da *Web* atual, que permitirá aos computadores e a humanos trabalhar em em cooperação. A *Web* Semântica interliga significados de palavras e tem como finalidade atribuir significado (sentido) aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo humano como pelo computador.

O projeto WS precisa usar uma outra linguagem que não HTML que separe o conteúdo da forma em que é exposto. A linguagem básica para isto é o XML, que permite inserir categorias semânticas nos dados que irão para a *Web*. A integração das linguagens ou tecnologias XML, *Resource Description Framework* (RDF), arquiteturas de metadados, ontologias, agentes computacionais, entre outras, favorecerá o aparecimento de serviços *Web* que garantam a interoperabilidade e cooperação. XML seria a linguagem de transporte de dados; RDF representaria a estrutura desses dados e a linguagem OWL, por exemplo, representaria a semântica desses dados, explicitando restrições sobre a semântica do mundo real. Aliadas a essas tecnologias, há também ferramentas para manipulação das ontologias (OntoEdit, Protegé, entre outras)

No contexto da *Web* Semântica, ontologia é um documento ou arquivo que formalmente define as relações entre termos e uma ontologia típica para a *Web* tem uma taxonomia e um conjunto de regras de inferência, conforme BERNERS-LEE, HANDLER E LASSILA, (2001).

Para que ocorra interoperabilidade entre máquinas e humanos são necessários dois requisitos básicos: (a) organização da composição de dados nos documentos que a compõem, formando um ambiente onde a informação esteja estruturada; e, (b) embutir semântica nestas informações, de maneira que os agentes, os softwares e os mecanismos de busca, tenham condições de compreendê-la. As principais propostas do projeto WS feitas pelo consórcio W3C são (MILLER, 2004):

- Construir uma representação de dados comum e de maneira estruturada que integre as mais variadas fontes informativas com o objetivo de se extrair conhecimento.
- Ampliar a abrangência do uso das informações mediante a ligação entre seus conceitos dentro de um contexto.
- Facilitar que as informações possam ser analisadas de maneira mais frutífera tanto por humanos como por máquinas, contribuindo para a descoberta do conhecimento.

Hoje, o fundamento da *Web* baseia-se no binômio *links*-recursos, que compõem o

conteúdo das páginas da grande rede. Entretanto é o usuário que determina se o que de fato ele lê, está de acordo com a informação que ele procura. Já na WS, é proposto aproveitar tais hiperligações, relacionando-as com os recursos e as camadas da informação, provendo contexto para os dados e facilidades para que as máquinas partilhem o conhecimento com os usuários.

Segundo (Daconta, 2003), a WS irá solucionar várias dificuldades das atuais tecnologias informativas, dentre elas:

- Informações não relevantes: Existe na *Web* muitas informações imprecisas, inúteis e irrelevantes disponibilizadas e tornam-se alvo dos motores de busca. Os filtros de busca enfrentam problemas em pesquisar páginas sem diferenciação dos assuntos, provocando falta de relevância no retorno aos usuários. Assim uma alternativa seria qualificar-se a informação, da própria informação.
- Integração dos conteúdos: Devido à vasta quantidade de informações oriundas de diversas fontes, ocorrem diferenças sintáticas, semânticas e estruturais, que inviabilizam o compartilhamento e a integração. Uma proposta para a solução deste conflito é o uso intensivo de metadados e ontologias, sob a perspectiva de desenvolver uma linguagem única, baseada na representação do conhecimento através de regras.
- Ausência de estruturação: A criação de recursos, documentos, arquivos, bancos de dados, etc, de maneira livre, por parte dos usuários da *Web*, é um enorme obstáculo a ser removido, no caminho da estruturação dos conteúdos. A falta de regras na elaboração dos recursos ocasiona prejuízos quando o usuário procura uma determinada informação, tornando o resultado da busca, muitas vezes ineficiente e sem relevância. Outras consequências do desprovimento de regras na *Web* são tempo excessivo na localização dos conteúdos, páginas não localizadas em virtude de uma mudança de URL (*Uniform Resource Locator*), retorno de uma quantidade gigantesca de informações, onde boa parte delas não atende ao que foi solicitado inicialmente, além dos problemas semânticos e das ambiguidades que frequentemente ocorrem. Para que os mecanismos de busca sejam realmente eficazes, as informações devem estar estruturadas e catalogadas segundo regras definidas, universalmente aceitas.

Assim, o propósito da WS é assistir às pessoas e não às máquinas, mas a sua exequibilidade depende diretamente da construção de instrumentos e mecanismos capazes de conferir lógica e semântica aos computadores.

1.2 Arquitetura da Web Semântica

Para que a WS se torne uma realidade, o W3C tem desenvolvido modelos que agrupam diversos tipos de tecnologia e linguagens. Ao longo dos anos, as pesquisas têm evoluído e modificado os agrupamentos como se pode ver pela figura 1:

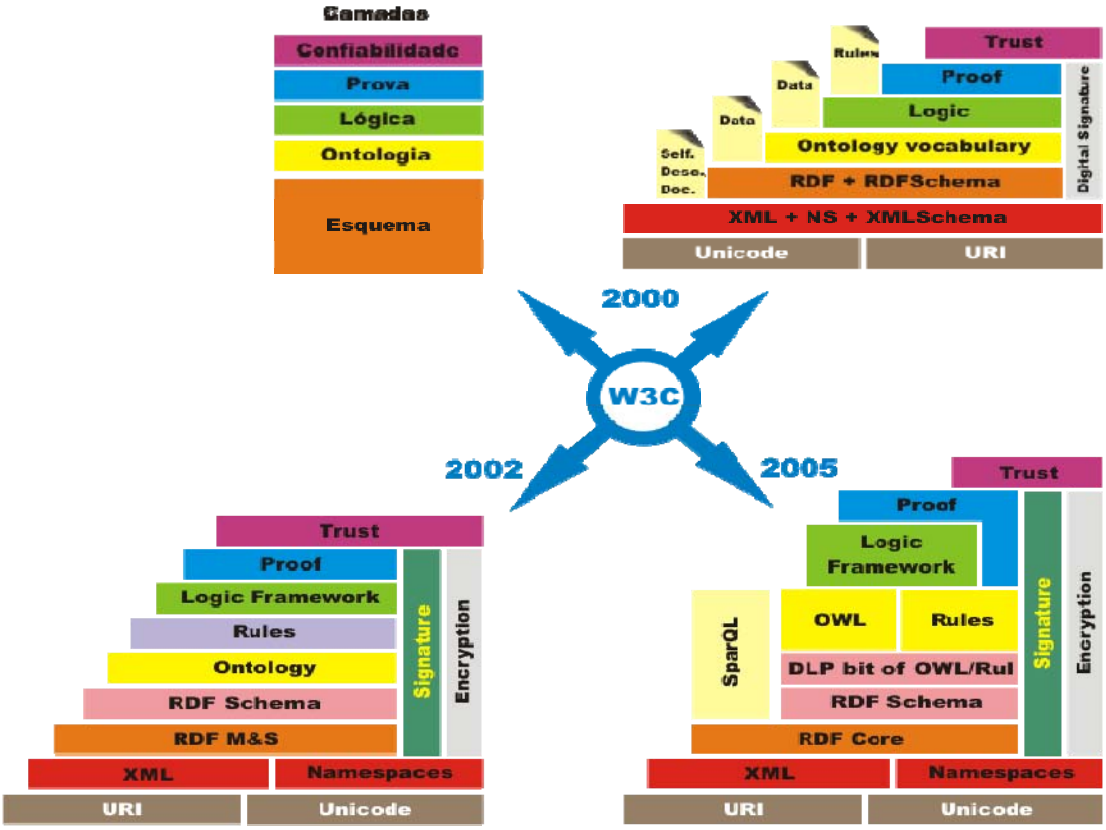

Figura 1 - Evolução das tecnologias da *Web* Semântica

Para uma melhor compreensão do envolvimento das linguagens e recursos no ambiente das pilhas semânticas, a arquitetura proposta pelo W3C será vista genericamente por camadas e em seguida item por item:

- Camada Estrutural: suas características principais são: i) possibilita a estruturação dos dados; ii) patrocina a unicidade, identificando os recursos univocamente; iii) trabalha a informação não somente provendo mecanismos para sua adequada representação, como também para sua armazenagem e intercambialidade, mantendo os dados de maneira íntegra e confidencial.

- Camada Sintática: duas importantes características desta camada são: i) associar o recurso ao dado estruturado e ii) verificar se o recurso está consistente em acordo às normas de sintaxe anteriormente descritas e validadas.
- Camada Semântica: é a geradora dos termos linguísticos que comporão a base de significados dos recursos. Nela também são definidos os relacionamentos entre estes termos. Trabalha com vocabulário formalizado e compartilhamento dos conceitos.
- Camada Lógica: é a que possibilita agrupar coleções de regras lógicas, com o objetivo de facilitar, sua execução através de agentes de software. Desta forma, estes podem operacionalizar automaticamente inferências e a verificação de coesão lógica dos recursos.
- Camada Prova: também chamada de camada de inferência. Seu objetivo é prover o acompanhamento das gerações das respostas, pela atuação dos agentes inteligentes de software. Para que assim aconteça, sua composição interna precisará estar escrita em linguagem que possibilite operar inferências.

Em seu núcleo criador de tecnologias, a WS se volta para a geração de anotação semântica de recursos e para a construção, evolução e integração de ontologias. O passo inicial para se atingir a WS é formatar os dados de maneira que os sistemas os compreendam nativamente. Um dos contextos mais expressivos de aplicação das tecnologias da *Web* Semântica é o de recuperação de informação na *Web*.

1.3 **Mecanismos de Busca**

A área de recuperação da informação, hoje, trata basicamente da recuperação de dados não estruturados, em especial de documentos textuais, em resposta a uma consulta do usuário (Greengrass, 2000). Com o crescimento da *Web* uma questão em aberto é como prover semântica e estrutura aos documentos armazenados.

O processo de recuperação da informação pode ser sintetizado em dois pontos principais: i) intersecção entre a pergunta formulada pelo usuário e a informação armazenada na Internet e ii) retorno relevante à pesquisa efetuada. Os mecanismos de recuperação são diversos e suas programações variam conforme a estrutura de seus algoritmos de busca e a indexação por eles realizadas. Os sistemas de recuperação da informação geralmente se baseiam em busca por palavra-chave ou busca por similaridade. Para isto, a maior parte

destes sistemas usa o modelo clássico ou o modelo estruturado. No modelo clássico, cada documento é descrito por um conjunto de palavras-chave – os termos de indexação, que buscam representar o assunto do documento e resumir seu conteúdo. Nos modelos estruturados, além das palavras-chave, são definidas algumas informações sobre a estrutura do texto, como as seções a serem pesquisadas e a proximidade das palavras. Dentre os modelos clássicos, temos o de Boole, o de Vetor, o de Probabilidade e o de Agrupamento.

O Modelo de Boole fundamenta-se na álgebra de mesmo nome, onde a recuperação da informação acontece, através do uso de expressões lógicas “e”, “ou” e “não”. Sua utilização apresenta vantagens, pois além de se obter respostas rápidas, este modelo também permite o armazenamento por índice, tendo custo mínimo.

O Modelo de Vetor entende o documento em si ou mesmo uma coleção deles como se fosse um campo vetorial. Assim cada palavra dentro da coleção, será representada no modelo por um vetor. O documento inteiro será um vetor de palavras, onde cada uma delas receberá um valor numérico, chamado grau de relevância (peso). Como cada termo possuirá um peso específico o vetor é genericamente representado como: [(termo_1, peso_1), (termo_2,peso_2),...,(termo_n,peso_n)]. O cálculo deste índice pode ser realizado de diversas formas e à quantidade de vezes que o vocábulo ocorre dentro do documento ou na coleção, dá-se o nome de frequência do termo no documento (FT). Os pesos também servem como parâmetro para investigar a similaridade entre documentos. Uma das maneiras para se obter o peso consiste em multiplicar sua frequência dentro do documento pelo log(N/nt), onde N é quantidade de termos em D (documento genérico) e nt é a quantidade de vezes que t se repete em D.

Utilizando um documento genérico D e um termo qualquer pertencente a este, chamado t, para se encontrar a frequência (FT) de t em D, procede-se ao seguinte cálculo:

FT = (quantidade de vezes que t aparece em D)

FID = Frequência Inversa do Documento

FID = log(N/nt) =>N é o quantidade de termos no documento e nt o quantidade de vezes que o termo t aparece no documento.

Para se calcular o peso de t, uma das maneiras é se multiplicar FT por FID a esta abordagem chamamos de TF-FID ou FT-IDF em inglês.

Este modelo de recuperação da informação apresenta algumas vantagens interessantes. Dentre elas o fato de retornar documentos que possuam graus de similaridade com a consulta inicialmente feita e melhoria do desempenho em função do sistema que aplica atribuição de pesos aos termos consultados. A principal desvantagem deste modelo é que

pode ocorrer que um documento extremamente relevante, não contenha nenhum dos termos pesquisados, desta forma o usuário não o receberá como retorno de sua consulta.

O Modelo de Probabilidade diferentemente do modelo anterior, este não utiliza pesos numéricos pré-definidos tanto para o índice de palavras do arquivo como para as consultas realizadas, nem tão pouco usa algum tipo de frequência na qual os termos aparecem no documento. A ordenação dos arquivos e o consequente peso das palavras é operado mediante um cálculo dinâmico, usando os vocábulos da consulta em relação aos documentos pesquisados e se fundamenta no princípio da ordenação probabilística (POP). Neste modelo existirão dois tipos de documentos: os relevantes para as consultas e os irrelevantes. Esta relevância será medida tendo como fundamento o número de consultas feitas pelos usuários. Aqui se tem o uso do princípio da retroalimentação. Uma grande vantagem deste modelo é permitir através do POP, uma ordenação ótima dos documentos, pois são classificados de maneira decrescente do ponto de vista de sua relevância.

O Modelo de agrupamento se baseia em *Clusters* são conjuntos disjuntos. Basicamente “clusterizar” significa que quaisquer documentos agrupados no mesmo cluster serão semelhantes entre si, possuindo aspectos parecidos. Já os agrupados em clusters diferentes não serão similares. Assim os documentos são classificados debaixo da ação de um algoritmo, não necessariamente agrupando-os pelos assuntos ou segmentos de conhecimento.

Os mecanismos de busca estão em constante desenvolvimento e a cada dia novos modelos e recursos são implementados, tendo como objetivo operar uma procura mais eficaz e consequentemente o devolver ao usuário respostas relevantes ao tema por ele pesquisado. Neste processo de evolução, já existem filtros de procura que dispensaram o uso da lógica tradicional (na qual para as informações eram dados ou o valor “0” ou o valor “1”) para trabalharem com a lógica nebulosa, que também atribui valores para as informações, só que entre “0” e “1”. Aplicações como estas, já usam graus de pertinência do termo pesquisado, calculando-o entre 0 e 1, o que amplia o universo da busca, pois permite verificar se o elemento indexado é de fato relevante e em que escala (valor numérico entre 0 e 1) esta relevância se apresenta.

Além da abordagem estatística, vista acima, os sistemas de recuperação da informação podem usar uma abordagem semântica. A abordagem semântica tenta implementar algum grau de análise sintática e semântica, visando compreender o texto em linguagem natural – presente nos documentos e nas consultas feitas pelo usuário, de forma análoga à do usuário humano. Boa parte dos aperfeiçoamentos operados nestes filtros de busca possui como base estreitar os laços com a WS. Alguns modelos pré-definidos voltados

para a WS e que integram os atuais mecanismos de busca são:

a. Modelo Especial Vetorial: Utiliza indexação automática de vocábulos-chave. O chamado elemento índice recebe um sistema de valoração por pesos, o qual numera o elemento, estabelecendo seu grau de importância dentro do documento. Pode ocorrer que dois vetores de elementos não possuam ortogonalidade entre si. Isto possibilita que ocorra a independência dos elementos índice, permitindo que conceitos por eles representados possam estar relacionados, o que pode ser verificado pelo aparecimento do mesmo elemento em documentos distintos. Ou ainda pelo estabelecimento de ligações semânticas entre os termos, através do uso de um tesauro.
b. Modelo Indexação Semântica: Questiona-se o uso de palavras-chave, para operar uma descrição mais aprofundada da pesquisa. O objetivo principal é compor uma união entre as pesquisas e os conceitos a elas relacionados nos documentos. Enquanto no modelo vetorial mapeiam-se vocábulos, que são os elementos índice, neste "busca-se mapear cada documento e cada consulta em um espaço menor, construído a partir dos conceitos relevantes que possuem os documentos no acervo" (SOUZA, 2006).
c. Modelo de Redes Neurais: Fundamenta seu funcionamento no processo cerebral humano e tenta reproduzir de forma artificial o que ocorre nas células nervosas, chamadas neurônios. Assim cada neurônio na rede (nó) representa um vocábulo ou um documento no acervo. Uma das principais características é a possibilidade de ocorrer o "aprendizado" do sistema, que se dá quando a rede consegue inferir quais os vocábulos que mais se assemelham ao que a pesquisa do usuário solicitou. É um modelo muito investigado atualmente, pois independente dos termos colocados na busca.
d. Modelo de Recuperação Textual: A recuperação da informação é realizada utilizando-se os componentes mais representativos do documento, como figuras, fotos, gráficos, quadros, tabelas e etc... Cria como ponto de conexão com a informação, a lembrança visual do usuário, que serve como ingrediente facilitador para que o sistema devolva a ele respostas mais eficazes.

Para Greengrass (2000), a grande maioria das soluções propostas para a recuperação da informação baseia-se na abordagem estatística; e mesmo as soluções semânticas apresentadas acabam por se basear fortemente nos métodos estatísticos.

Nesta dissertação, a recuperação da informação é realizada por consulta SQL e via programação em PHP, utilizando a função strpos, disponível nas versões PHP 4 e PHP 5.

Esta função encontra a posição da primeira ocorrência de uma string, realizando uma busca booleana simples.

Na próxima seção, são apresentados os feeds.

1.4 *Feeds*

Os programas chamados agregadores diferenciam-se dos mecanismos de busca, pois enviam ao usuário constantemente informações atualizadas sobre diversos assuntos. Estes programas são também conhecidos como readers, ou leitores de feeds. Os *Feeds* baseiam-se em RSS (*Really Simple Sindication*), que é um subconjunto da linguagem XML. Através dos *Feeds* RSS, os usuários podem receber as atualizações dos canais de informação nos quais se cadastraram, sem ter que visitá-los. Estes aplicativos são muito usados em blogs e em sites de notícias como G1 e Terra, por exemplo. O agregador pode concentrar informações recebidas de diferentes canais de notícias agrupando-as no computador do cliente. Existem alguns que funcionam junto ao navegador não necessitando de instalação na máquina do usuário. As principais características dos Feeds são:

- Seu uso está voltado para que o cliente *Web* receba atualizações de notícias, novidades sobre artigos ou blogs, sem precisar procurá-las em um mecanismo de busca ou mesmo visitando o site fonte. Tais informações poderão ser lidas por ele em seu próprio computador, mediante o programa agregador (leitor de *Feeds*).
- O processo de assinar ou mesmo o de cancelar o recebimento dos *Feeds* é extremamente simples, sendo suficiente a ação de adicionar o link ao programa agregador (para assinar o *Feed*) ou então retirá-lo, para cancelar a sua vinda. Não é necessário o envio de mensagens eletrônicas para o distribuidor dos *Feeds*.
- A ausência da necessidade de cadastrar o e-mail do interessado em receber os *Feeds* preserva-o de ser alvo de spams e mensagens maliciosas voltadas para o roubo de senhas e a infestação por malwares e vírus.
- Programas agregadores mais robustos, além de trazer as informações ao usuário, permitem que ele as organize de forma personalizada, classificando-as conforme desejar.

Um dos programas agregadores usados é o *FeedReader*, que possui código aberto e pode ser configurado para o idioma português. Este leitor de *Feeds* permite buscas internas

nas notícias que recebe sendo de fácil instalação e manuseio. Sua tela de notícias é apresentada na figura 2, com as principais funcionalidades, marcadas pelas letras A,B,C e D.

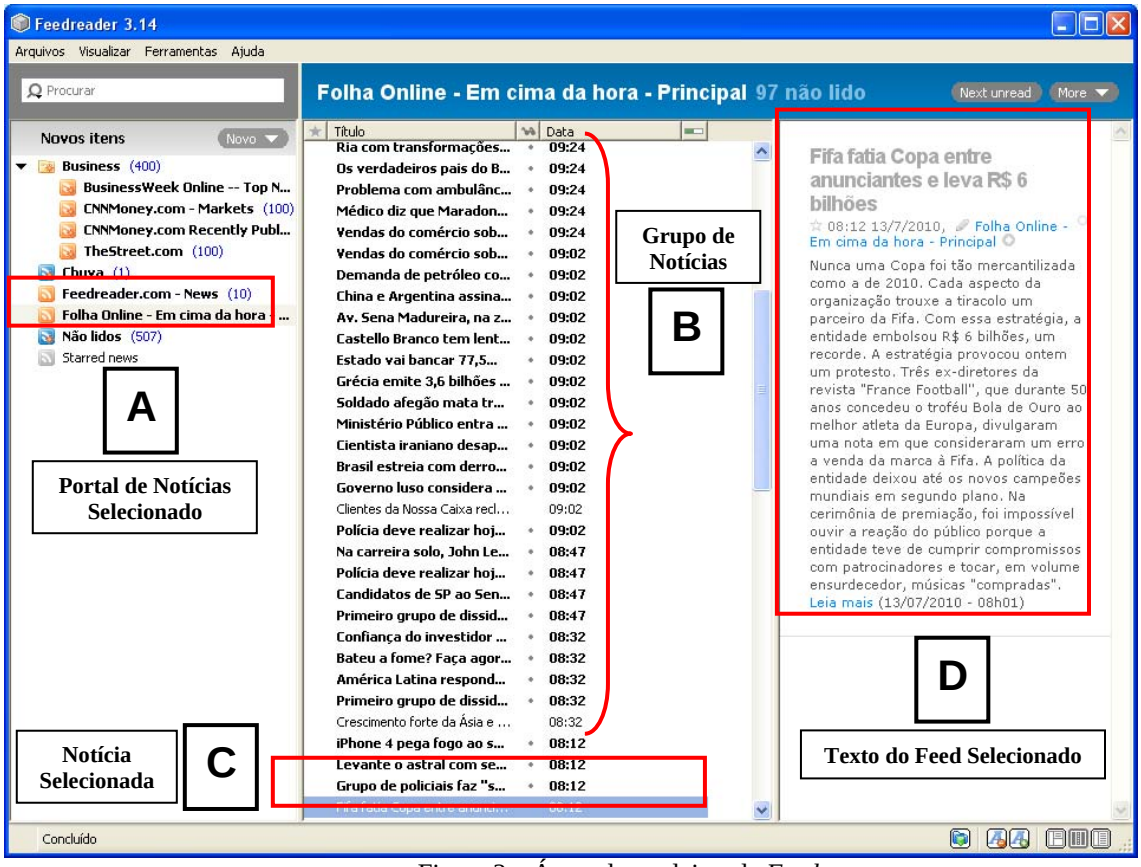

Figura 2 – Áreas de um leitor de *Feeds*

O serviço de *Feed* pode ser visto como um avanço na forma de recuperar informação para o usuário, mas como mencionado no capítulo introdutório desta dissertação, ainda que opere com classificações semi-automáticas de notícias, ele não contempla a semântica das notícias. A proposta deste trabalho é estender as funcionalidades atuais do serviço de *Feed*, para dotar a recuperação/classificação da informação de uma semântica. Para tanto, é proposto o uso de ontologias.

1.5 **Ontologias**

Para (Gruber, 1993), ontologias são especificações apresentadas de maneira formal (baseadas no conhecimento do mundo real) estabelecidas sob conceitos compartilhados. (Guarino, 1998) as define como uma teoria lógica que aponta para o real significado dos termos componentes de um vocabulário formal. Em (Pereira, 2007, p.110), assim são explicados os termos mais relacionados às ontologias e às suas definições:

> Especificação formal refere-se ao fato de uma ontologia poder ser lida pela máquina. Explícita, significa que são usados conceitos e que as restrições no seu uso são explicitamente definidas. Conceitualização, diz-se de um modelo abstrato de algum fenômeno no mundo, sendo identificados os conceitos relevantes do fenômeno. Partilhada, reflete a noção de que uma ontologia captura conhecimento consensual, isto é, que não é privado de um indivíduo, mas sim aceito por um grupo.

Para (Berners-Lee; Handler; Lassila, 2001), o cerne da WS é fazer com que o conteúdo disponível na Internet possua semântica explícita e que esta seja processável pelos computadores em todo o mundo. As ontologias desempenham importante papel na busca por este objetivo, pois conseguem representar de maneira adequada, as visões e os processos semânticos necessários para que as máquinas "compreendam" o que os usuários (humanos) desejam encontrar, realizar e obter, no universo *Web*. Uma visão esquemática das características das ontologias é apresentada na figura 3.

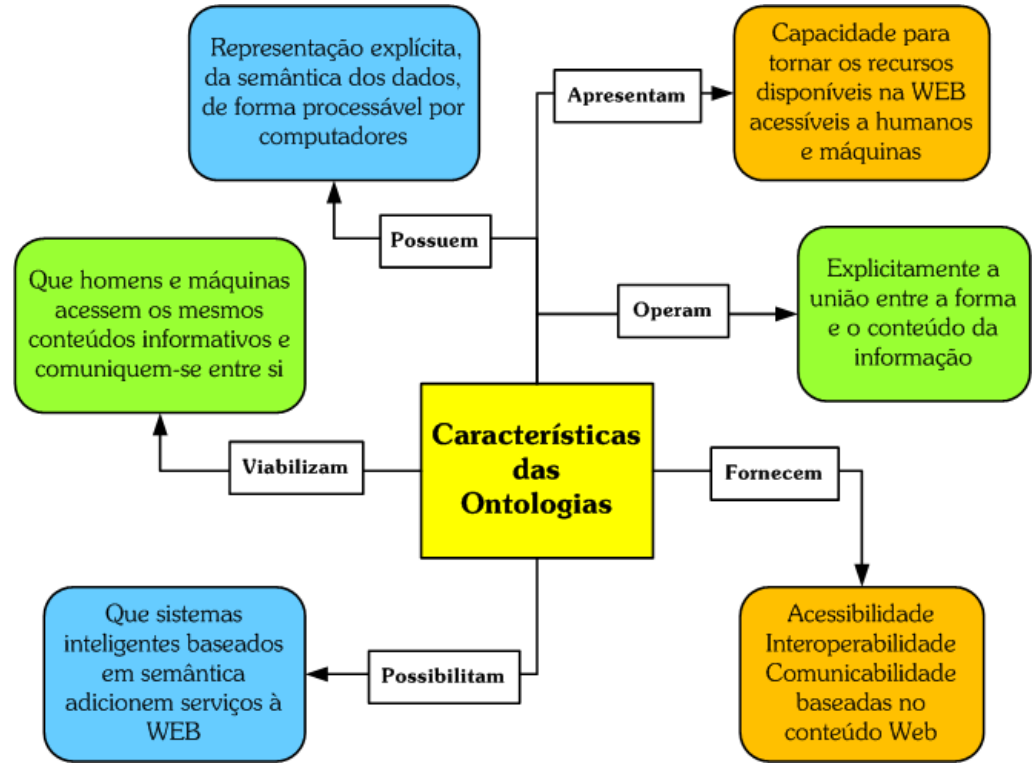

Figura 3 – Características das ontologias

Para Noy e McGuiness (2001), os pontos principais para se iniciar a montagem de uma ontologia são: especificar tanto o domínio como o alvo da ontologia; pesquisar a possibilidade de se fazer o reuso de ontologias já construídas; criar uma lista que possua os termos mais importantes; estabelecer a definição para cada classe e de suas respectivas

propriedades; e, construir uma hierarquia, que envolva as classes já definidas. Estas autoras apresentam três regras básicas para nortear o processo de desenvolvimento e manutenção de uma ontologia:

- Há diversas formas para se realizar a construção de um determinado domínio, não havendo apenas uma maneira correta de modelá-lo. A melhor opção para fazê-lo, está ligada à subjetividade dos alvos que se pretende alcançar e à capacidade de poder prever possíveis adições ao conjunto;
- O processo de construção é iterativo, ou seja, repete-se reiteradas vezes. Obtendo-se uma versão preliminar da ontologia, a aplicação entrará em uma dinâmica de evolução, (que inclui revisões ao trabalho inicial) e tenderá a ser infinita, ou mesmo continuar ocorrendo enquanto o ciclo de vida da ontologia perdurar;
- Os conceitos formadores da ontologia devem ser reflexos tanto para os objetos que os representam como para os relacionamentos existentes no domínio que se estuda e que se pretende modelar. Tais premissas são necessárias para que de fato a ontologia consiga reproduzir situações verdadeiras do mundo real.

Ainda de acordo com Noy e McGuiness (2001), o processo de desenvolver uma ontologia contempla dois tipos de definição: a de classes (com a organização destas em uma taxonomia, através de uma hierarquia composta por subclasses e superclasses) e a das propriedades delas, com a consequente descrição dos valores permitidos (construção de restrições). Além disto, há o desdobramento final, que inclui o preenchimento de valores (compatíveis) para as propriedades instanciadas. São elementos de uma ontologia:

- Conceitos - São as classes da ontologia e representam objetos do mundo real, especificados dentro do domínio estudado.
- Propriedades - Atributos das classes. Poderão sofrer restrições quanto aos seus conteúdos.
- Relacionamentos - Ligações que se estabelecem entre as classes.
- Axiomas - É o conjunto de condições necessariamente verdadeiras, que devem ser obedecidas, para que tanto a interpretação dos conceitos como o uso de seus relacionamentos possa ser validado. Possuem aspectos delimitadores e restritivos.

- Instâncias - Indivíduos componentes de uma ontologia. São as materializações provenientes dos conceitos e seus relacionamentos.

Gruber (1993) e Pérez (1999) também descrevem alguns requisitos importantes que devem constar na montagem de ontologias, que são:

- O vocabulário usado (termos componentes da ontologia) deve ser descrito de forma clara e objetiva, apoiando-se em uma documentação expressa em linguagem natural.
- Mostrar coerência entre os termos e suas definições, com a intenção de prover um grau suficiente de confiabilidade para geração de inferências.
- Ter a condição de agrupar novos termos (extensibilidade), de maneira que não se desfaça a estrutura já montada.
- Tanto a codificação quanto o compromisso ontológico devem ser mínimos. Para isto, a codificação deve fornecer condições de ser interpretada e entendida sem que haja a intervenção de tecnologias adicionais particulares. Já no compromisso ontológico, ser mínimo significa possibilitar que a ontologia construída seja compartilhada e reusada.
- Apresentar disjunção de classes para preservar a identidade conceitual. Maximizar a padronização dos nomes dos termos utilizados. Agrupar em subclasses os conceitos que forem semelhantes e de forma correspondente ao que ocorre na engenharia de software, buscar a integração da ontologia, com outras ontologias, através do uso de módulos na sua construção.

A seguir serão descritos os elementos que ampliam a capacidade semântica das ontologias:

- Dicionários – São os agrupamentos de palavras que compõem determinado idioma. Neles as palavras são referenciadas por ordem alfabética, apresentam suas significações e podem trazer a parte fonética a elas acoplado, como também aspectos mórficos e sintáticos. O seu principal objetivo é esclarecer o que significa cada um de seus vocábulos, de modo que todo um grupo de indivíduos fale e entenda do mesma modo. Assim, pode-se considerar que o dicionário mapeia os termos utilizados em uma determinada língua definindo-os, de maneira que não haja dúvida quanto à sua interpretação por ocasião de seu uso entre as pessoas.

- Glossários – São especializações do dicionário, que abrangem um grupo específico de palavras ou expressões, geralmente são termos de uso mais técnico ou regional. Definem fatos, ocorrências, situações dentro de um domínio específico do conhecimento.
- Índices – Assim como glossários e dicionários, os índices também são listas de palavras dispostas em ordem alfabética, podendo de acordo com a necessidade serem apresentados em ordens numéricas e alfa-numéricas. São constituídos pelos termos mais relevantes existentes em um determinado documento mapeados pela posição em que eles ocorrem. Alguns exemplos típicos do uso de índices são: a) índice remissivo em um livro, onde os conceitos e vocábulos mais importantes são listados e mostrados em junção com o número da página onde estão escritos; b) catálogo de publicações de uma biblioteca onde o índice associa a obra consultada, a uma prateleira e estante específicas, localizando univocamente o livro, revista ou artigo pretendido.
- Taxonomias – O vocábulo provém do grego (taxis, divisão, organização, para classificar e nomos, lei, administração) e se constitui na ciência que procura classificar tanto seres vivos, objetos inanimados, conceitos e etc. Praticamente todas as coisas podem ser classificadas sob a ótica de algum tipo de taxonomia. A taxonomia se distingue da ontologia porque esta se preocupa em exaurir completamente um determinado domínio extraindo-lhe todos os conceitos, aquela se detém em aplicar sobre os conceitos uma hierarquia, subdividindo-os em classes e subclasses, mostrando inclusive seus relacionamentos.
- Tesaurus – De acordo com Librelotto, Ramalho e Henriques (2005, p.8) o “tesaurus é um instrumento que reúne termos escolhidos a partir de uma estrutura conceitual previamente estabelecida, destinados à indexação e à recuperação de documentos e informações em um determinado domínio.” Ele se distingue do dicionário, pois este aplica definições a vocábulos dos mais variados domínios do conhecimento, enquanto que um tesaurus se limitará a descrever expressões que estejam interligadas entre si e façam parte do domínio já pré-definido. Embora ele também seja um tipo de índice, possui a característica de encorpar e aprimorar as taxonomias tornando-as mais robustas quando da descrição de algum domínio em particular, uma vez que permite relacionamentos enre as classes que ultrapassam a simples união hierárquica.
- Redes Semânticas – São estruturas de representação do conhecimento, onde os objetos são os nós de um sistema de grafos. Os relacionamentos existentes entre os objetos

tomam a forma de arcos (relações binárias) que os interligam. Os nós são estruturados em uma taxonomia. Um dos objetivos pretendidos pela Inteligência Artificial ao usar as redes semânticas, é prover a representação do conhecimento baseado na organização de conceitos (Librelotto, Ramalho e Henriques, 2005). Para Gonçalves (2007), elas são um caso de particularização das ontologias. A rede de palavras em inglês (Wordnet) é um exemplo de rede semântica. É uma ontologia linguística que agrupa, em seu arcabouço, sinônimos, adjetivos, substantivos e verbos. Este modelo tem como um dos objetivos principais operar a desambiguação dos termos, através da execução de algoritmos que realizam ações sintático-semânticas sobre os textos.

1.6 **Vantagens do Uso das Ontologias**

Uma das premissas fundamentais para que qualquer mecanismo de busca e representação do conhecimento funcione de maneira eficaz é que exista um conteúdo de dados e informações bem organizado e definido formalmente. As ontologias trabalham sobre este formalismo conceitual, podendo assim representar a informação semântica e semi-estruturada. Através disto, elas executam a aquisição, a manutenção e o reuso do conhecimento. Entre as principais vantagens para o uso das ontologias se destacam as seguintes:

- O vocabulário fornecido pela ontologia está alicerçado em bases conceituais sólidas. Isto permite que ela represente adequadamente o conhecimento do que ocorre no mundo real, sem ambiguidade de interpretação.
- As ontologias operam o conhecimento compartilhado e colaborativo de um determinado domínio, fundamentando-se nas definições formais dos conceitos envolvidos.
- Proporcionam a reutilização do conhecimento.
- Podem ser expressas em outros idiomas.
- Operam inferências nos conteúdos informacionais gerando conhecimento.
- Servem como apoio à engenharia de sofware e atuam na interoperabilidade dos sistemas.
- Com base no formalismo dos conceitos, ontologias diferentes podem se integrar, compondo uma outra maior e mais abrangente.

- Existem vários softwares onde elas podem ser desenvolvidas, editadas e testadas por motores de inferência, que conseguem verificar sua consistência.
- São usadas por sistemas gerais de banco de dados.
- Facilitam o acesso inteligente a caudais de dados, onde ocorrem informações textuais e semi-estruturadas.
- Proporcionam que os filtros de busca retornem aos usuários *Web*, respostas mais relevantes e eficientes às suas pesquisas.

1.7 **Construção de Ontologias**

A complexidade do trabalho que envolve a construção de uma ontologia é facilitada pelo uso de metodologias e softwares, com os quais elas podem ser elaboradas, construídas, testadas e editadas.

1.8 **Metodologias para Construção de Ontologias**

A construção de uma ontologia deve ser realizada como um projeto de criação de software, sendo executadas etapas como levantamento de requisitos, (que exige que especialistas do domínio sejam consultados), implementação (não somente ocorrendo o uso de uma linguagem, mas também com uma metodologia de desenvolvimento), testes de eficiência e consistência do modelo, com a consequente observação dos resultados. Além disto, são extremamente desejáveis que itens como reuso, extensibilidade e interoperabilidade entre os sistemas sejam contemplados durante a montagem da estrutura ontológica. O trabalho em si deve ser de fácil manuseio e manutenção dos itens já integrados, permitindo alterações, acréscimos e finalmente a evolução da ontologia elaborada. Embora já existam alguns ambientes computacionais que atuem como facilitadores para o desenvolvimento de ontologias, sua construção continua sendo uma tarefa complexa e que demanda tempo e custos. A engenharia de ontologias ainda dá seus primeiros passos e atualmente não há uma metodologia que seja considerada amplamente completa para ser usada. Na literatura pesquisada foram selecionados alguns métodos que se apresentam como os mais usados pelos pesquisadores e desenvolvedores no contexto da ciência da computação e da engenharia e gestão do conhecimento.

1.8.1 Metodologia Uschold & King

Este método foi elaborado para fornecer as bases para a construção da ontologia Enterprise, que se propunha a modelar as atividades empresariais. Após sua composição, ela tornou-se em uma coleção de termos e definições relevantes de uso comum nas empresas privadas. Foi desenvolvida no Projeto *Enterprise Applications* pelo Instituto de Inteligência Artificial da Universidade de Edimburgo, tendo como empresas associadas: IBM, *Lloyd's Register, a Logica UK Limited, e Unilever*. Os passos que englobam esta metodologia são:

- Condição de Existência: O porquê de se montar uma ontologia (no caso a empresarial)
- Construção da Ontologia: nesta etapa serão identificadas as palavras-chave (conceitos) e os relacionamentos existentes no domínio que se deseja mapear. Tanto conceitos como os relacionamentos pertencentes ao domínio serão definidos textualmente de maneira formal e sem ambiguidades, que é na prática a codificação da ontologia.
- Paralelamente aos dois itens já mencionados, deve-se operar questionamentos sobre a reutilização de ontologias que já existam e abarquem o domínio em uso ou partes dele.
- Avaliação da Ontologia: Utilizando atruições técnicas entre elas a verificação de requisitos (uso de especialistas na área), processa-se a validação das questões de competência sempre comparando-as com o que ocorre no mundo real.
- Documentação: Os documentos referentes à ontologia deverão conter toda a descrição do processo de elaboração. Este item é o mais importante quando da análise sobre o possível reúso da ontologia.

1.8.2 Metodologia Methontology

Desenvolvida pelos professores Gómez-Pérez, Fernándes e Juristo no Instituto Politécnico de Madrid, possui a característica de apresentar todo ciclo de vida de uma ontologia. Conforme Breitman (2005), seu objetivo é ser um instrumento facilitador na integração do grupo que trabalha para operar a construção da ontologia subdividindo-o em três equipes: i) aqueles que ficarão responsáveis pelas atividades de gerenciamento das ontologias; ii) os que exercerão responsabilidade com os processos referentes ao seu desenvolvimento e iii) os encarregados de realizar o suporte. Conforme Fernández, Gómez-Pérez e Juristo (1997), a *Methontology* possui os seguintes itens mostrados na figura 4 :

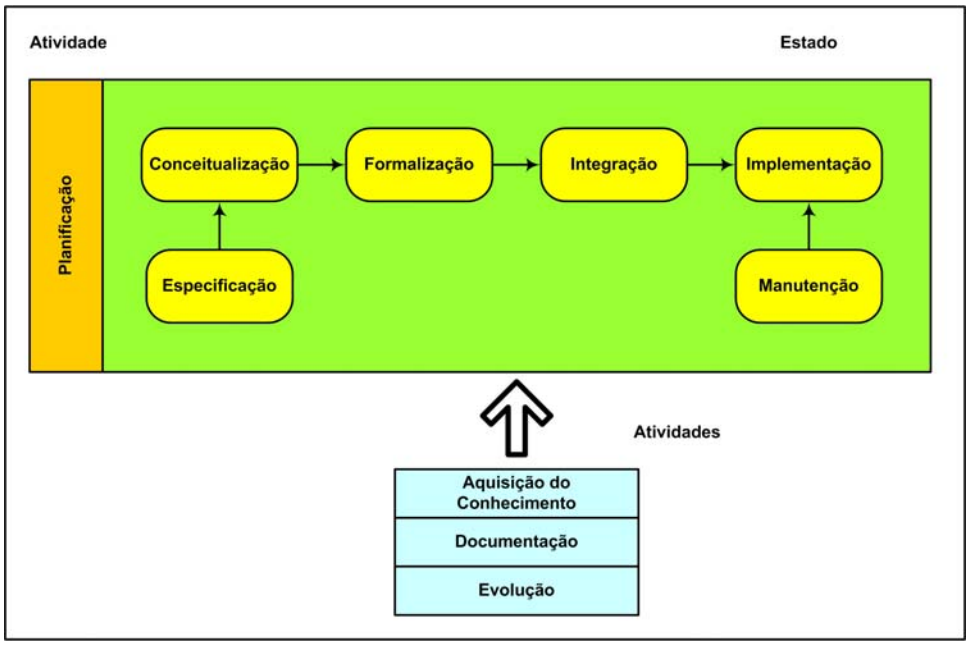

Figura 4 – Etapas da Metodologia Methontology

Esta metodologia delineia o processo de construção subdividindo-o em tipos de atividades para serem executadas. Brandão e Lucena (2002) agrupam as realizações do método em três vertentes de atividades: i) as de gerenciamento; ii) as orientadas ao desenvolvimento e iii) as de suporte que são realizadas em paralelo com as que ocorrem no item ii, e as desdobram em doze etapas explicadas a seguir:

1. Atividades que envolvem o gerenciamento do projeto

- Planejamento: As tarefas que serão realizadas são identificadas, organizadas e estimadas em quanto de recursos e de tempo consumirão para serem executadas. É essencial cumprir este ponto se o desejo da equipe é operar o reuso de ontologias.
- Controle: Garante que todo o planejamento apresentado no item anterior seja de fato cumprido.
- Qualidade: Assegura que os produtos finais que serão apresentados pela equipe, (ontologias, softwares e documentações), sejam eficientes e satisfaçam aos objetivos iniciais do projeto.

Figura 14 – Tela de Inclusão de Fontes Emissoras de RSS

2. Atividades que envolvem o desenvolvimento

- Especificação: Esta etapa localiza a ontologia no tempo e no espaço, provendo-lhe suas bases iniciais (o porquê de sua construção), a finalidade que terá (para quê será construída – seu futuro uso) e finalmente exporá quem a usará após ser montada (seus destinatários finais).
- Conceitualização: Esta fase se caracteriza pelo levantamento e seleção dos conceitos envolvidos no domínio da ontologia que se pretende construir. “Para se realizar este levantamento poderão ser utilizadas as técnicas empregadas na engenharia de requisitos” ROQUE (2009).
- Formalização: Esta passo é responsável pela transformação dos conceitos obtidos no item anterior em um modelo formalizado e que seja semi-computacional.
- Implementação: Tem como finalidade determinar uma linguagem de programação e codificar nela os modelos já concebidos.
- Manutenção: Envolve alterações no projeto original, com o objetivo de efetuar ajustes, correções e evoluções na ontologia.

3. Atividades que envolvem o suporte (paralelas ao desenvolvimento)

- Aquisição do conhecimento: “Atividades de aquisição do conhecimento sobre um determinado domínio”, BRANDÃO E LUCENA (2002).
- Avaliação: Os produtos em fase de geração (ontologias, ambientes, softwares e as documentações) receberão o crivo técnico dos especialistas das respectivas áreas, através de frames de referência.
- Integração: Esta fase se volta para a execução de atividades essenciais que tornarão possível operar reuso de ontologias existentes. O método em questão prevê que conceitos, possam ser utilizados em outros processos ontológicos.
- Documentação: Neste item, todos os passos feitos para a construção da ontologia, serão exaustivamente detalhados. Isto contribuirá para futuras alterações na ontologia, garantindo sua constante evolução.