Web Semântica: Conceitos Básicos e Tecnologias Associadas[1]

Tatiane Domingos Dias
Neide Santos
Departamento de Informática e Ciência da Computação
Instituto de Matemática e Estatística - Universidade do Estado do Rio de Janeiro
neide@ime.uerj.br

***Resumo**: O tutorial tem como objetivo apresentar os principais conceitos e tecnologias sobre a Web Semântica. O trabalho aborda os problemas da estruturação, entrega e interoperabilidade de informações na Web, as ontologias e as tecnologias propostas para representação do conhecimento e de seu conteúdo semântico.*

1. Introdução

Uma área atual de pesquisa e desenvolvimento em Ciência da Computação trata da questão da semântica envolvida na recuperação da informação na Web. A Web Semântica objetiva dar uma estrutura aos conteúdos das páginas Web, criando um ambiente onde agentes de software perambulam pelas páginas para desempenhar tarefas sofisticadas requisitadas pelos usuários. Entre estas tarefas, está a busca contextualizada da informação.

Um dos objetivos originais da Web era a troca de informação entre pessoas, mas de forma de que os computadores pudessem participar da comunicação, ajudando os usuários. Os computadores na Web, atualmente, têm papel somente no direcionamento e entrega de informações, não tendo acesso ao conteúdo das páginas, porque essa informação está estruturada para utilização pelas pessoas e não por máquinas. Hoje, temos uma Web de documentos e não de informações. Por isso, os computadores oferecem ajuda limitada no acesso e processamento da informação, deixando as funções de extração e interpretação dessa informação a cargo dos usuários.

A Web Semântica visa resolver este problema, estruturando o conteúdo das páginas Web de forma que a informação possa ser interpretada pelas máquinas. A proposta não é a de uma Web separada da atual, mas uma extensão da mesma, baseada em documentos – as ontologias - descrevendo relacionamentos entre objetos e contendo informação semântica dos mesmos para automatizar o processamento pelas máquinas.

Na Web há uma quantidade imensa de informações não pertinentes que é fornecida pelos processos de busca. As ferramentas de busca enfrentam a dificuldade de executar pesquisas entre documentos que não estão diferenciados em termos de assunto, qualidade e relevância. A tecnologia atual não é capaz de diferenciar uma informação comercial de uma educacional, ou informação entre idiomas, culturas e mídia. É necessário haver informações de qualificação da própria informação, chamada de metadados, para ser possível classificá-las e tornar os processos de busca mais eficazes. Algumas dessas novas estruturas necessárias já foram definidas e outras ainda estão sendo desenvolvidas pelo *Word Wide Web Consortium* (W3C). W3C é um composto de organizações interessadas na definição e desenvolvimento de novos conceitos, protocolos e padrões de estruturas para a Web, visando obter maior eficácia de seus recursos.

O trabalho desenvolvido pelo W3C tem como foco o acesso universal à Web Semântica para desenvolver um ambiente onde a informação seja expressa de maneira a possibilitar a automatização de tarefas e a melhor utilização dos recursos por parte dos usuários. Outro objetivo é a criação de uma Web confiável, oferecendo confiabilidade e possibilitando que as pessoas assumam a autoria e responsabilidade por suas publicações. Um dos princípios fundamentais utilizados no *design* de tecnologias para a Web á a interoperabilidade. As especificações de linguagens e protocolos para Web devem ser compatíveis entre si de forma a permitir que qualquer tipo de hardware ou *software* utilizado para acessar a Web possa trabalhar em conjunto com estas especificações. Para tanto, o W3C faz uso de princípios como interoperabilidade, evolução e descentralização para desempenhar suas tarefas de identificação de novas tecnologias para a Web e de projeção e padronização das mesmas.

O objetivo deste tutorial é apresentar os principais conceitos e tecnologias sobre a Web Semântica, pois a idéia de uma Web com semântica é recente, e muitos dos conceitos e tecnologias envolvidos estão ou dispersos ou mal estruturados e divulgados

[1]Tutorial extraído de Dias, Tatiane D,. Web Semântica: Fundamentos e Tecnologias. 2001. Trabalho de Conclusão de Curso (Graduação em Informática) - Universidade do Estado do Rio de Janeiro.

2. Evolução Histórica da Web

Originalmente, o computador era visto somente como hardware. Na década de 80, ele se transformou em um sistema capaz de simular jogos, processar textos e elaborar apresentações. Hoje em dia, tornou-se um portal para uma rede de troca de informações e transações comerciais. Como conseqüência, as tecnologias que dão acesso a essas informações textuais, não estruturadas e heterogêneas se tornaram tão essenciais quanto às linguagens de programação nas décadas de 60 e 70. A Internet, mais especificamente a tecnologia Web, deu início a estas mudanças e acarretou uma série de transformações de caráter tecnológico, social e econômico. A Web passou a propiciar uma nova plataforma para o desenvolvimento de aplicações com acesso distribuído por diferentes partes do planeta. Antes de seu surgimento, os principais serviços utilizados na Internet eram a transferência de arquivos, o correio eletrônico e a emulação de terminal, e restritos aos meios acadêmicos e militares. O uso generalizado da Internet só veio a acontecer, em 1992, com o surgimento da Web, que organizou as informações na Internet por meio de hipertexto e, em um segundo momento, tornou a interação do usuário com a rede mundial mais amigável.

Inicialmente, a Web era um projeto desenvolvido, a partir de março de 1989, por Tim Berners-Lee no CERN (Laboratório Europeu para Física de Partículas), para acessar informações estanques espalhadas pelos diversos laboratórios na Europa, tendo evoluído para um serviço usado globalmente. O que era um sistema baseado em buscas por hipertexto teve seu crescimento viabilizado pelo traço cooperativo da Internet, ou seja, pela colaboração mútua entre os componentes da rede. A aparente simplicidade da Web gera obstáculos para seu próprio desenvolvimento, já que a tecnologia utilizada atualmente limita a manipulação da informação.

O primeiro objetivo do projeto da Web era criar um ambiente em que pudéssemos trabalhar melhor em grupo tanto no trabalho quanto em casa. A idéia era que criando uma web de hipertexto, os grupos de usuários seriam forçados a utilizar um vocabulário comum entre eles para que não ocorressem mal entendidos e, em algum momento, teriam um modelo na web dos planos e idéias em discussão no grupo. O precursor da web foi um programa para uso próprio, chamado “Enquire”, desenvolvido por Tim Berners-Lee, em 1980, quando ele ainda trabalhava no CERN (Laboratório Europeu de Física de Partículas). Este programa tinha o propósito de manter registros da complexa rede de relacionamentos entre pessoas, programas, máquinas e idéias espalhadas pelos diversos laboratórios na Europa. Mais tarde, em 1989, ele viria a apresentar uma proposta para a Web que, na verdade, era uma extensão deste programa pessoal.

Muito freqüentemente desperdiçamos tempo e esforço tentando registrar em documentos idéias e definições discutidas e firmadas em reuniões ou encontros de grupos e acabamos por causar mal entendidos por causa da subjetividade de interpretação de cada pessoa. A Web foi desenhada para ser utilizada como um instrumento de prevenção de mal entendidos. Para que isso funcione, a Web não tem que ser apenas fácil de se navegar, mas também auto-explicativa. Qualquer informação disponível na Web pode ser facilmente assimilada e qualquer informação que esteja faltando pode ser facilmente adicionada. A Web deve ser um meio de comunicação entre as pessoas; comunicação através do compartilhamento de conhecimento. Isso requer que computadores, redes, sistemas operacionais e programas sejam transparentes aos usuários, disponibilizando somente uma interface intuitiva e o mais direta possível com a informação.

O segundo objetivo da Web, dependente do primeiro, é baseado na premissa que se há informação disponível na Web então é possível estruturarmos esta informação, criando um mapa de relacionamentos e dependências. Isso possibilitaria o acesso dos programas a estas informações e permitiria que eles nos ajudassem em sua análise e gerenciamento. A estruturação do conteúdo semântico da informação das páginas web criaria um ambiente, onde agentes de *software* executam tarefas solicitadas pelos usuários e pessoas e computadores possam trabalhar em cooperação, deixando a cargo dos computadores qualquer tarefa que possa ser reduzida a um processo racional.

Apesar de inicialmente estar direcionada ao trabalho em grupo, a Web se desenvolveu rapidamente como um ambiente de compartilhamento de documentos e não de informação que pudesse ser utilizada pelos computadores. Isso ocorreu devido à facilidade de publicação de documentos na Web e um ambiente onde poucos publicam e milhares utilizam. Ainda são poucos os que publicam porque o mercado de *software*s de edição de páginas ainda está amadurecendo lentamente. A falta de editores de fácil utilização não é o único empecilho para a consolidação da Web como um ambiente de colaboração. Há também a necessidade de ferramentas que forneçam controle de acesso confiável garantindo que somente pessoas autorizadas tenham acesso às informações, e que estas ferramentas sejam de fácil manipulação tornando transparente a seus usuários os detalhes pertinentes aos sistemas operacionais.

Na verdade, há também um limite do que é possível de ser feito somente pelos humanos, sem interferência das máquinas. Uma das maiores queixas dos navegadores iniciantes é a quantidade imensa de informações não pertinentes fornecida pelos processos de busca na Web.

Algumas das novas estruturas necessárias já foram definidas e outras ainda estão sendo desenvolvidas pelo

W3C. Seus interesses estão voltados para novas áreas e conceitos que estão emergindo como intranet, comércio eletrônico e *e-learning*, e que podem agregar a evolução da Web. Eles procuram alcançar consenso sobre protocolos a serem aplicados nestas áreas, regras que possibilitem a comunicação entre máquinas, pois é a partir da criação desses protocolos que se torna hábil o desenvolvimento de novas aplicações capazes, então, de se comunicarem. Essa é a chave para qualquer desenvolvimento na Web e para a criação de um ambiente realmente interativo.

O trabalho desenvolvido pelo W3C tem como foco os seguintes objetivos [W3C01] [W3C02]:

- Acesso universal: colaborar para que a Web se torne acessível a todos a partir do desenvolvimento e utilização de tecnologias que contemplem as grandes divergências culturais, educacionais, de recursos, e principalmente as limitações físicas dos usuários em todo o mundo;
- Web Semântica: desenvolver um ambiente onde a informação seja expressa de maneira a possibilitar a automatização de tarefas e melhor utilização dos recursos por parte dos usuários;
- Web confiável: guiar o desenvolvimento na Web considerando cuidadosamente os aspectos legais, comerciais e sociais da tecnologia em questão. Criar uma Web que ofereça confiabilidade e possibilite que as pessoas assumam a autoria e responsabilidade por suas publicações.

W3C concentra seus esforços em três principais tarefas [W3C01] [W3C02]:

- Visão: promover e desenvolver sua visão a respeito do futuro da World Wide Web. Devido à contribuição de milhares de pesquisadores e engenheiros que trabalham em organizações filiadas ao W3C e a comunidade da Web, é possível que o W3C identifique os requerimentos técnicos necessários para que a Web se torne um verdadeiro espaço universal de compartilhamento de informação;
- Design: projetar tecnologias para a concretização de sua visão tendo como base três princípios fundamentais: interoperabilidade, evolução e descentralização. Esses princípios serão descritos mais adiante;
- Padronização: contribuir para reforçar a padronização de tecnologias Web produzindo especificações, denominadas recomendações, que descrevem as etapas de construção. Estas recomendações estão disponíveis para serem acessadas por qualquer pessoa interessada e sem nenhum custo.

Como foi especificado anteriormente, existem três princípios fundamentais utilizados no design de tecnologias para a Web [W3C01] [W3C02]:

- Interoperabilidade: as especificações ou recomendações de linguagens e protocolos para Web devem ser compatíveis entre eles e permitirem que qualquer tipo de hardware ou *software* utilizado para acessar a Web possa trabalhar em conjunto com estas especificações;
- Evolução: a Web precisa ser capaz de acomodar tecnologias futuras, e para isso, conceitos como simplicidade, modularidade, compatibilidade e extensibilidade, devem ser considerados na especificação de tecnologias e protocolos. Assim, as chances de compatibilidade das tecnologias dispostas atualmente na Web com tecnologias emergentes aumentam muito;
- Descentralização: este é o princípio utilizado pelos sistemas distribuídos e o mais difícil de ser considerado no design de tecnologias. É necessário eliminar o máximo de dependências existentes em centrais de registro, gerando um ambiente flexível e fundamental para a evolução não só da Web, mas da Internet como um todo.

Para atingir os objetivos de criação de uma Web de acesso universal e que contenha informações estruturadas de maneira a serem utilizadas pelas máquinas na automação de tarefas e informações confiáveis em que possam ser identificados os autores e responsáveis por suas publicações, o W3C faz uso de princípios como interoperabilidade, evolução e descentralização.

3. Web Semântica

O primeiro passo para dotar a Web de semântica é a construção das chamadas ontologias de domínio. Para Berners-Lee, Hendler e Lassila [TBL01], uma ontologia típica para a Web é composta de uma taxonomia e um conjunto de regras de inferência. Mas elas não seriam suficientes para imprimir semântica à Web, requerendo a adoção de tecnologias novas, como por exemplo, XML (*Extensible Markup Language*) [XML] e RDF (*Resource Description Framework*) [RDF]. XML possibilita a criação de *tags*, campos de texto que ficam escondidos nas páginas web. Os programas ou scripts podem fazer uso dos *tags* de várias formas, mas o programador precisa saber o significado de cada *tag* criado pelos autores das páginas para utilizá-los. Ou seja, XML permite que o usuário adicione estruturas arbitrárias a seus documentos, mas não permite representar o significado de cada estrutura. Este seria o papel desempenhado pelo RDF - expressar significado às estruturas. O RDF codifica os *tags* em um conjunto de triplas, sendo cada tripla dotada de um sujeito, verbo e objeto de uma sentença simples. Essas triplas podem ser escritas utilizando XML *tags*. Em RDF, um documento pode fazer assertivas sobre relações entre coisas tais como Maria (sujeito) é irmã (verbo) de Pedro (objeto). Essa estrutura tende a ser uma maneira natural de descrever a maioria das informações processadas pelos computadores. O sujeito e o objeto desta sentença são identificados, cada um, por um indicador universal denominado URI (*Universal Resource Identifier*), como

os utilizados em *links* nas páginas web, já que a URL (*Uniform Resource Locator*) é o tipo mais comum de URI. Os verbos também seriam identificados por URIs, facilitando a definição de novos verbos ou conceitos apenas pela criação de novas URIs em qualquer lugar na Web [TBL01].

Utilizando URIs para codificar informações de relacionamentos entre objetos assegura-se que esses conceitos não são somente palavras escritas em um documento, mas também são definições únicas acessíveis a todos na Web. Se, por exemplo, tivéssemos acesso a vários bancos de dados contendo informações sobre pessoas, inclusive seus endereços e se quiséssemos encontrar alguém que reside em um código de endereçamento postal específico, precisaríamos saber que campo em cada banco de dados se refere ao nome desta pessoa e qual se refere ao código, para realizarmos esta busca. O RDF seria capaz de representar esta informação através de sentenças que utilizam URI para cada termo.

É possível que vários bancos de dados utilizem identificadores diferentes para conceitos iguais. Um programa que queira comparar ou utilizar informações de distintos bancos de dados precisa saber que diferentes termos têm o mesmo significado. Esse objetivo é alcançado através da criação de coleções de informações denominadas Ontologias. Ontologia, na filosofia, significa teoria a respeito da natureza da existência. Pesquisadores e estudiosos das áreas de Inteligência Artificial e Web incluíram esse termo em seus jargões com um significado adaptado que é documento ou arquivo que define formalmente as relações entre os termos.

A taxonomia define classes de objetos e relacionamentos entre os mesmos. Por exemplo, um endereço pode ser definido como um tipo de localização e códigos de cidade podem ser definidos como aplicáveis somente à localizações. Classes, subclasses e relacionamentos entre entidades são muito úteis para uso na Web. Entre elas existe o conceito de herança de propriedades, ou seja, é possível associarmos propriedades às classes que suas subclasses herdam automaticamente essas propriedades. Por exemplo, se códigos de cidade são definidos como do tipo cidade que, por conseguinte, possui Web sites, então podemos associar um determinado código de cidade a um site Web sem existir um relacionamento direto entre os dois.

As regras de inferência são de essencial importância para as ontologias. Através delas é possível expressarmos, por exemplo, que "se um código de cidade estiver associado a um determinado estado, então os endereços que utilizam este código de cidade também estão associados a este estado". Um programa poderia deduzir que se a rua Paissandu, localizada na cidade do Rio de Janeiro, pertence ao estado do Rio de Janeiro e, por conseguinte, ao país Brasil, então, as informações devem seguir os padrões brasileiros de formatação. O computador realmente não entende esse tipo de informação, mas consegue manipulá-lo de maneira a desempenhar um papel mais significante e eficaz de ajuda ao usuário.

Páginas Web baseadas em ontologias são o começo de muitas soluções para os problemas de terminologia. O significado de alguns termos e códigos XML utilizados nas páginas podem ser definidos através da criação de ponteiros para ontologias. Continuarão existindo alguns problemas inerentes aos usuários, pois se uma pessoa cria um ponteiro para uma ontologia que define um endereço através da informação de cep e outra pessoa cria um ponteiro para uma ontologia que também define endereço, mas utilizando a informação de caixa postal é necessário que ambas as ontologias ou outro serviço web qualquer seja capaz de identificar que a informação de cep é equivalente a de caixa postal.

As ontologias podem agregar valor ao funcionamento da Web, já que podem ter várias aplicações diferenciadas. A forma mais simples seria aumentar a precisão dos mecanismos de busca de informação. Os programas de busca pesquisariam somente em páginas que fizessem referência a um conceito pré-definido ao invés de pesquisar todas as que contenham palavras-chave. As aplicações mais avançadas as utilizariam com o objetivo de relacionar o conteúdo das páginas às suas estruturas existentes de conhecimento e regras de inferência.

O potencial da Web Semântica será realmente compreendido quando forem desenvolvidos programas que sejam capazes de efetuar buscas de informação de diferentes fontes disponíveis na Web, as processem e compartilhem os resultados com outros programas. A eficácia dos programas, baseados em agentes, tende a aumentar na medida em que houver mais conteúdo na Web estruturado de maneira que possa ser utilizado pelos computadores. Os agentes seriam responsáveis por captar as necessidades do usuário, pesquisar e disponibilizar os resultados esperados de forma interativa.

3.1. Heterogeneidade da Informação

A integração de informações na Web é um assunto muito discutido pelos estudiosos da área. A variedade de fontes de informação distintas com diferenças sintáticas, semânticas e estruturais entre elas é muito grande, tornando o compartilhamento, integração e resolução de conflitos entre essas informações um problema de difícil solução.

Outra questão a ser tratada seria a criação ou remoção de fontes de informação, o que teria que ser realizada com extrema cautela de forma a não causar grandes impactos ao ambiente integrado. Deve-se considerar que as fontes de informação podem ter

capacidades computacionais diferentes, podendo variar desde sistemas de banco de dados a arquivos. As informações podem variar de não estruturadas, como imagens e vídeos, a semiestruturadas, como arquivos de *e-mail* e páginas Web.

A heterogeneidade estrutural e semântica da informação na Web, atualmente, é imensa e a maioria das propostas de integração ainda adota soluções com alto índice de centralização, tornando seu uso na Web inviável. Para tratar esses problemas é necessário considerar questões relevantes como a utilização de metadados e ontologias, visando a busca de uma linguagem única, capaz de estruturar e representar conhecimento e regras.

3.2. Busca e Recuperação da Informação

Um dos motivos do grande sucesso da Web é sua liberdade de publicação de informação. Encontra-se facilidade para criação de páginas Web e não é necessário, por exemplo, pedir autorização de qualquer pessoa para criar *links* entre páginas, nem mesmo do próprio criador da página.

Devido a isso, existe uma enorme quantidade de documentos e recursos de todo tipo disseminado na Web, tais como: bancos de dados, artigos, programas, arquivos, etc. Por serem criados de forma autônoma, sem preocupação com regras de estruturação, catalogação e descrições de suas propriedades, essas informações são difíceis de serem abrangidas pelos mecanismos de pesquisa, ocasionando demora e ineficácia na localização de informações. Alguns problemas enfrentados pelos mecanismos de busca e recuperação de informações são: demora na localização de informações; informações não localizadas devido às mudanças de URLs; recuperação de um número elevado de informações que, em sua maioria, não atendem às expectativas dos usuários; e, recuperação de informações fora do contexto solicitado pelo usuário devido a problemas de semântica e ambigüidade.

Devido a esses problemas, a busca pelo aprimoramento das ferramentas e mecanismos de busca direcionados à localização e recuperação das informações é um tópico importante e um grande desafio. A efetividade desses mecanismos de busca depende principalmente da maneira pela qual as informações foram estruturadas e catalogadas na Web. Documentos podem ser estruturados e organizados de várias formas diferentes na Web e as ferramentas de busca têm que utilizar mecanismos de recuperação adequados para cada tipo de organização.

As ferramentas de busca estão classificadas em quatro categorias que estão, sucintamente, descritas a seguir [HAB] [KAM] [UTM]:

- Pesquisa em diretórios: essas ferramentas efetuam pesquisa por tema e de forma hierárquica. Começam as buscas a partir de um tópico genérico, ramificando em subtópicos específicos. Disponibilizam a informação em forma de diretórios e o próprio usuário tem que navegar na árvore de diretórios a procura de informações mais específicas a respeito do tema pesquisado. Estas ferramentas são mais eficazes para pesquisas de temas amplos.
- Máquinas de busca: efetuam pesquisa através de palavras-chave. Utilizam bancos de dados que são constituídos de palavras-chave e URLs [URL] que foram previamente pesquisadas nas páginas web e copiadas para o banco de dados por robôs (*crawlers*). A pesquisa é feita no banco de dados e fornece como resultado uma relação de URLs de páginas Web onde o usuário pode encontrar algo sobre o tema pesquisado. Por ser pesquisa textual, muitas vezes, os resultados não correspondem às expectativas do usuário.
- Diretórios com máquinas de busca: elas utilizam tanto a pesquisa em diretórios quanto por palavras-chave. Na parte referente à pesquisa em diretórios, ela segue um percurso hierárquico, desde assuntos genéricos aos mais específicos e, em cada pausa ao longo deste percurso, disponibiliza-se uma máquina de busca permitindo que o usuário efetue uma pesquisa por palavra-chave dentro daquele universo de diretórios. Não é indicada para pesquisas complexas e difíceis devido ao problema de imprecisão.
- Meta Busca (múltiplos mecanismos de busca): utilizam recursos de várias máquinas de busca em paralelo e é conduzida através de palavras-chave. O resultado é apresentado na forma de uma lista de informações obtida de acordo com cada mecanismo de busca envolvido ou de forma integrada.

A Web Semântica visa tornar a Web um ambiente de acesso inteligente à informação heterogênea e distribuída através de agentes de *software* que utilizarão mecanismos de busca mais acurados para disponibilizar informações aos usuários. A heterogeneidade da informação dificulta a integração de conteúdo na Web. Agora veremos como é possível a descrição de forma homogênea da informação através do uso de metadados.

3.3. Metadados na Web

Metadados, também conhecidos como informações sobre dados, são utilizados para documentar e organizar de forma estruturada e padronizada as informações de documentos com o objetivo de facilitar e tornar mais efetiva a busca e recuperação da informação na Web. O metadado é estruturado com elementos de descrição do conteúdo dos dados. Cada bloco de informações deve conter, por exemplo, autor, título, data de publicação etc. e para cada campo pode conter as seguintes informações: nome do campo, descrição do campo, tipo de dados, formato, etc, e qualquer informação que seja relevante para a recuperação da informação. No

contexto da Web, três aspectos devem ser considerados no desenvolvimento de metadados: descrição de recursos, produção e uso de metadados [IAW].

O primeiro aspecto refere-se a quais informações estarão sendo consideradas nos metadados. Um metadado tem que ser suficientemente flexível para capturar informações de diversas fontes distintas. O segundo aspecto refere-se à construção de metadados. Os metadados nada mais são do que sumários sobre uma determinada informação. Utilizar trabalho humano para gerar estes metadados seria caro e cansativo. A tendência é automatizar este processo o máximo possível. Já o terceiro e último aspecto trata de como os metadados serão acessados e utilizados. Eles têm que estar disponibilizados de maneira que possam ser processados preservando seu conteúdo semântico. Quanto à sua utilização, podem servir de forma especialmente relevante na localização de recursos na Web, contendo informação descritiva dos recursos e onde estes podem ser encontrados.

No entanto, devido ao aspecto dinâmico dos recursos na Web, a disponibilização de metadados causa alguns desafios, já que freqüentemente novas versões de recursos são acrescentadas a Web e documentos são renomeados e disponibilizados em outros endereços (URL) [URL]. Outras questões também importantes a serem discutidas a respeito dos metadados são [IAW]:

- Possibilidade de descrever um recurso a partir de mais de um conjunto de qualificadores devido ao grande número de padrão de metadados;
- Necessidade de existência de um conjunto de padrões específicos para cada tipo de recurso de forma a acomodar todos os tipos diferentes;
- Internacionalização dos padrões, já que a maioria dos padrões é baseada em qualificadores em Inglês;
- Metadados devem ser gerados na medida em que um recurso é criado e disponibilizado na Web, sendo alterado na medida em que o recurso é modificado. Entretanto alguns tipos de metadados mais específicos podem ser gerados à parte, tais como: críticas sobre um filme ou artigos;
- Metadados também são dados e por isso apresentam características de armazenamento e acesso, e dificuldades de interpretação de seu conteúdo.

Padrão de Metadados

A criação de um único padrão de metadados que aborde todas as áreas do conhecimento humano é um assunto muito discutido e de expectativa remota, já que existem muitos problemas a serem solucionados primeiramente como a necessidade de um padrão composto de inúmeros qualificadores para que seja possível abranger os diversos domínios existentes. Isso torna a catalogação exaustiva e exige um conhecimento mais específico devido aos vários domínios de conhecimento. Mas é possível estabelecer padrões de metadados de forma que as organizações possam ser convidadas e encorajadas a utilizá-los no sentido de contribuir para a documentação de suas informações. O esforço neste sentido deve ser conjunto para que haja uma padronização e uma divisão das tarefas. Fortemente associados aos metadados existem determinados padrões que podem ser adotados.

Neste tutorial, consideramos apenas os padrões utilizados para descrição dos recursos na Web, e que são utilizados para recuperação da informação. Este tipo de padrão de metadados apresenta uma forma estruturada a partir de um conjunto de qualificadores simples e genéricos que objetivam a descoberta e gerenciamento dos recursos. Dentre os padrões que se encontram nesta categoria estão o IAFA (*Internet Anonymous FTP Archive*) [IAFA], SOIF (*Summary Object Interchange Format*) [SOI] e Dublin Core [WKL] [DC]. Dentre os mais utilizados, encontra-se o padrão Dublin Core.

Padrão Dublin Core

O padrão Dublin Metadata Core Element Set [DKL] [DC], ou Dublin Core, foi desenvolvido pelo W3C com a finalidade de contemplar os seguintes objetivos: simplicidade de criação e manutenção; semântica de fácil compreensão; interação com padrões já existentes ou emergentes; escopo e aplicabilidade internacional; capacidade de extensão; e, interoperabilidade entre coleções e sistemas de indexação.

Dublin Core tem como objetivo catalogar e classificar os documentos eletrônicos (textos, mapas, imagens) de forma a facilitar a recuperação dos mesmos na Web. É um dos padrões mais utilizados devido sua facilidade de manipulação e extensa capacidade de descrição dos recursos. É constituído de 15 elementos qualificadores, que possuem as seguintes propriedades:

- *Name*: nome único de identificação do qualificador;
- *Label*: nome como o qualificador é conhecido;
- *Definition*: Descrição que representa o conceito e natureza do qualificador;
- *Comment*: Informação adicional a respeito do qualificador (opcional);
- *See Also*: *Link* para maiores informações sobre o qualificador (opcional)

Os elementos qualificadores de Dublin Core são:

Title: Título do objeto
Creator: Pessoas responsáveis pelo conteúdo do objeto
Subject: Tópico abordado pelo objeto
Description: Descrição textual do conteúdo do objeto
Publisher: Entidade responsável pela disponibilização do objeto
Contributor: Pessoa ou organização que contribui intelectualmente na criação do objeto
Date: Data da criação ou publicação do recurso
Type: Forma como o conteúdo é expresso
Format: Formato em que o objeto é disponibilizado (HTML, DOC, PDF, etc).

Identifier: Identificador único do objeto
Source: Informação sobre as fontes de informação que contribuíram para a criação do conteúdo do objeto
Language: Idioma
Relation: Relacionamentos com outros objetos
Coverage: Características temporais e espaciais
Rights: Informações sobre os direitos autorais do objeto

Veremos a seguir como as ontologias se integram ao conceito da Web Semântica.

3.4. Ontologias

Na Web Semântica, a ontologia é utilizada no contexto de compartilhamento do conhecimento e tem como objetivo a especificação explícita e formal de uma conceituação. Assim, Ontologia é a descrição explícita e precisa de conceitos e relações que existam em certo domínio de conhecimento [GRU]. Uma Ontologia requer o uso de um vocabulário específico para descrever os requisitos para um determinado domínio e também um conjunto de axiomas lógicos necessários para imprimir semântica ao significado pretendido pelas palavras do vocabulário. Assim a Ontologia pode gerar um ambiente com informações documentadas, confiáveis, e de fácil manutenção e reutilização.

Existem duas principais propriedades das ontologias que devem ser analisadas devido a sua importância no processo de criação das mesmas [BEZ]. São elas:

- Compartilhamento: refere-se a capacidade de compartilhar informações comuns entre sistemas. Diferentes sistemas devem utilizar as mesmas ontologias de modo a ter as mesmas definições de conceitos, minimizando assim a ocorrência de várias ontologias para conceituação das mesmas informações; e,
- Filtragem: definição do que realmente é relevante a ser extraído de um determinado sistema utilizando modelos de abstração que levam em consideração somente parte da realidade, deixando de lado características indesejáveis da informação.

Através dessas duas propriedades, uma ontologia deve ser capaz de extrair informações de modo a criar um modelo de sistema enxuto, significativo e integrado. Mas também é preciso que uma ontologia seja flexível o bastante para aceitar informações de diferentes naturezas.

Geralmente, em uma ontologia, existem três níveis ou tipos de informação em uma ontologia:

- Terminológica: constituído de um conjunto básico de conceitos e relações da ontologia. Normalmente conhecida como a camada de definição;
- Assertiva: conhecida como camada de axiomas da ontologia, é constituída de um conjunto de assertivas aplicáveis aos conceitos e relações; e,
- Pragmática: denominada camada de caixa de ferramentas. Constitui-se de informações técnicas a respeito de conceitos e relações da camada terminológica, informações estas que não podem estar classificadas em nenhum das outras duas camadas. Nesta camada podemos encontrar informações, por exemplo, a respeito da forma como um determinado conceito ou relação é apresentado ao usuário.

As ontologias são de grande importância para a Web Semântica, pois conseguem embutir significado, sem ambigüidade, às informações através da criação de vocabulários, interconexões semânticas entre os termos e regras de inferência e lógica sobre um determinado domínio de conhecimento, facilitando a interpretação e recuperação da informação por agentes de *software* e viabilizando também o intercâmbio de informações entre eles. Possibilitam também um mecanismo de pesquisa mais apurado e restrito às informações realmente relevantes, automação de tarefas que exijam raciocínio, e permitem que os agentes atuem como guias, sugerindo opções e caminhos e auxiliando o usuário no alcance de seus objetivos.

Os agentes desempenham papel importante na Web Semântica. Agentes são programas que capturam o conteúdo de várias fontes na Web, processam estas informações e fazem intercâmbio desses resultados com outros programas. Possuem um certo grau de autonomia e são capazes de realizar tarefas que auxiliem o usuário no desempenho de suas atividades, de acordo com seus interesses [JHD].

Por esses motivos, a pesquisa na área de agentes é considerada um caminho promissor para o desenvolvimento de aplicações para a Web, em especial aquelas relacionadas a sistemas distribuídos e inteligentes [WOOL]. Dada à existência de inúmeras pesquisas nessa área, há muitas definições sobre o significado de agentes, entre elas: “um agente é um sistema computacional encapsulado, que está situado em algum ambiente e é capaz de executar ações flexíveis e autônomas no ambiente de forma a alcançar seus objetivos” [WOOL]. A idéia de ambiente é utilizada de forma genérica, podendo se referir a qualquer meio físico ou lógico, composto de aspectos heterogêneos ou não. Já por ações autônomas que um agente pode executar, compreende-se qualquer ação que possa ser realizada sem intervenção humana, e flexíveis, no sentido de não contemplarem somente ações pré-determinadas, ou seja, que possuam uma tabela de ocorrências possíveis em uma ambiente. Esses tipos de agentes são denominados agentes inteligentes.

A flexibilidade dos agentes inteligentes implica na consideração de três características:

- Reatividade: capacidade de perceber o ambiente onde atuam e responder em tempo satisfatório às mudanças que ocorrem;

- Pró-atividade: possuir um comportamento centrado no alcance da meta e assim sendo capaz de tomar a iniciativa de certas tarefas; e,
- Habilidade social**:** capacidade de interagir com o ambiente, com outros agentes e com os usuários.

Os agentes podem ser classificados em diversas categorias. Algumas delas consideram a mobilidade do agente, outras os classificam de acordo com seu sistema de raciocínio, dividindo-os em reativos e deliberativos (ou cognitivos). Outros classificam os agentes segundo o grau de autonomia, aprendizado e cooperação que possuem. Atualmente, pode-se encontrar na literatura uma grande variedade de aplicações que, de formas distintas, fazem uso do conceito de agentes para implementar algumas funcionalidades, mas o que fica realmente evidente é o papel que o agente desempenha. Os agentes estão assim classificados da seguinte forma [NWA]:Agentes de Colaboração; Agentes de Interface; Agentes de Informação; Agentes Móveis; e, Agentes Híbridos. O tipo de agente utilizado na Web Semântica é o agente de informação, responsável pela pesquisa e recuperação de conteúdo na Web.

Um agente de informação deve ser capaz de se adaptar a seus inúmeros usuários e ao conteúdo a ser disponibilizado, e possuir uma única interface para acesso a múltiplos repositórios de informação. Além disso, ele deve ser capaz de localizar, recuperar e integrar informações, e, ainda, procurar por informações de forma pró-ativa em fontes distribuídas, evitando intervenções do usuário sempre que possível. E, mais importante, o agente só deve disponibilizar informações que sejam realmente do interesse do usuário.

Esses agentes, capacitados com as qualidades já descritas, podem solucionar problemas atualmente encontrados em sistemas de recuperação da informação, tais como: necessidade de soluções integradas, recuperação de informação distribuída, expansão de termos para refinamento de busca, interfaces e navegação, filtragem, recuperação eficiente, identificação de preferências do usuário.

3.5. Tecnologias para Representação da Informação

A Web Semântica não está relacionada apenas ao formato do conteúdo de um recurso, mas também à forma como este conteúdo será disponibilizado e interagirá com outros recursos na Web. Para que a Web Semântica funcione de forma efetiva é necessário que as informações estejam estruturadas disponibilizadas de tal maneira que possibilite a implementação de um raciocínio automatizado por parte das máquinas.

Os metadados e as ontologias são conceitos importantes, pois permitem a criação de coleções estruturadas de informações e conjuntos de regras de inferência, estabelecendo assim um domínio de conhecimento com vocabulário comum e informação semântica a respeito de seu conteúdo e suas relações. Entretanto, há inúmeros domínios de conhecimento distintos a serem representados a partir de diferentes padrões de metadados e Ontologias e para isso são necessárias arquiteturas de alto nível, capazes de prover suporte à codificação e intercâmbio dessa variedade de metadados desenvolvidos de forma independente na Web. E sempre com o objetivo de estar contemplando a interoperabilidade tanto semântica quanto sintática e estrutural.

Nesta seção do tutorial, serão vistas as principais tecnologias que despontam atualmente para o desenvolvimento da Web Semântica e suporte a interoperabilidade de informação.

A linguagem criada para prover sintaxe à informação é XML. Considerada a mais importante e capaz de codificar todo tipo de informação de forma que esta possa ser transferida entre recursos na Web alcançando assim a interoperabilidade sintática. A semântica é conseguida através da criação de Ontologias específicas para cada domínio de conhecimento utilizando linguagens como SHOE [SHOE], XOL [XOL], OIL [OIL], DAML [DAML], entre outras. Já a interoperabilidade sintática é responsabilidade do RDF, modelo capaz de prover uma estrutura padrão para a informação [TBL01].

Estes padrões foram desenvolvidos pelo consórcio W3C [W3C] no intuito de que, quando utilizados em conjunto, sejam capazes de fornecerem informações estruturadas, passíveis de serem processadas pelas máquinas e intercambiadas entre recursos de forma mais inteligente.

Extensible Markup Language – XML

Originalmente desenvolvida com o objetivo de dar suporte a larga escala de *softwares* de editoração eletrônica que estavam surgindo no mercado, a linguagem XML é atualmente uma ferramenta de muita importância para o intercâmbio entre recursos da grande variedade de informações disponíveis na Web.

Alguns dos objetivos que o consórcio W3C visava alcançar quando disponibilizou a primeira versão da linguagem, em Fevereiro de 1998, são [XML]:

- Possibilitar a internacionalização da mídia independente da editoração eletrônica;
- Permitir que as indústrias definam protocolos de plataformas independentes para o intercâmbio de informações recursos, especialmente as pertinentes ao comércio eletrônico;
- Disponibilizar informações aos *softwares* agentes de forma a permitir o processamento automático pelas máquinas;
- Facilitar o desenvolvimento de *softwares* especializados na manipulação de informações distribuídas em várias fontes na Web;

- Facilitar o processamento das informações pelos usuários através de *software*s de custo mais acessível;
- Permitir que os usuários disponibilizem as informações da forma desejada, sem estarem presos a controles de estilos de formatação;
- Facilitar o fornecimento de metadados e possibilitando uma pesquisa e recuperação da informação mais eficaz e satisfatória ao usuário.

Hoje em dia, a maioria dos documentos disponibilizados na Web utiliza o sistema de marcação provido pela linguagem HTML. XML supre alguns objetivos que HTML não conseguiu alcançar e que foram identificados ao longo do percurso como pontos muito importantes para esse tipo de linguagem. A linguagem XML é similar a HTML em alguns aspectos. XML também faz uso de marcações, denominadas *tags*, mas estes têm a finalidade somente de delimitar e descrever parte das informações, deixando a interpretação a cargo das aplicações que as utilizam. Além disso, os documentos em XML podem ser utilizados por mais de uma aplicação ao mesmo tempo, já que estas acessam somente os *tags* que são relevantes para elas e fazem sua própria interpretação dos mesmos.

O fator que realmente diferencia XML de outras linguagens deste mesmo tipo é sua capacidade extensiva, pois provem um formato de dados para estruturação de documentos sem utilização de um vocabulário específico. Isso permite que a XML seja identificada como uma linguagem de aplicabilidade universal, já que é possível criar ilimitados *tags* para inúmeros tipos de documentos. XML consegue que a formatação do documento seja tratada separadamente de sua estrutura, que pode ser descrita com maior riqueza de informações, pois é passível de ser personalizada pelo autor.

A entidade principal na linguagem XML é o *element*. Este é constituído normalmente de dois *tags*, denominados *tag* inicializador e *tag* finalizador, e do texto delimitado entre eles. Geralmente os *tags* são representados, respectivamente, como <pessoa> e </pessoa>. Um elemento pode conter outro elemento ou texto. Se um elemento não possuir nenhum conteúdo, ele pode ser abreviado para <pessoa/>.

Os elementos possuem conceitos de parentesco, ou seja, quando são criados vários elementos aninhados, é preciso finalizar os elementos mais inferiores, denominados filhos, primeiro para, por último, finalizar o elemento raiz. Todo documento XML deve possuir um elemento raiz e este deve necessariamente ser finalizado ao final do documento. É possível associar aos elementos atributos com valores. Um atributo é codificado como um par, "palavra=valor", dentro de um *tag* do elemento. Quando um documento em XML possui um *tag* raiz, os *tags* estão corretamente aninhados e os atributos são únicos diz-se que é um documento ou bem formado, sendo possível organizá-lo em uma estrutura de árvore.

A utilização de XML não implica na obtenção de interpretação específica do conteúdo de um documento. Um documento XML é constituído de entidades, sub-entidades e valores, compondo assim uma árvore ordenada e valorada, mas sem nenhum tipo de semântica. É viável codificar qualquer tipo de estrutura de dados em sintaxe ambígua, mas a linguagem XML não especifica a semântica e a forma de utilização da informação dentro do contexto do documento. Então os recursos responsáveis pelo intercâmbio destas informações precisam assegurar vocabulário comum (nome de elementos e atributos), como será sua utilização e seu significado. Então veremos a seguir dois mecanismos de especificação de vocabulário a ser utilizado em documentos: DTD e XML Schema.

Document Type Definition (DTD) e XML Schema

DTD e XML Schema são mecanismos utilizados para especificar a estrutura de documentos escritos em linguagem XML. Então é possível e recomendado verificar se um documento está elaborado conforme as regras de estrutura especificadas em um DTD ou XML Schema com a finalidade de determinar se este documento é válido ou não [XML]. DTD e XML Schema são conhecidos também como informação de cabeçalho de documentos XML e são responsáveis por:

- Descrever regras estruturais que os *tags* devem seguir no documento, tais como, se é permitido utilizar elementos aninhados, atributos necessários e seus valores possíveis, e estruturada de nome de elementos e atributos;
- Lugares no documento onde é permitida a utilização de texto normal;
- Listar os recursos externos ou entidades externas utilizadas no documento;
- Declarar os recursos internos ou entidades internas que podem ser requeridas no documento; e,
- Relacionar os tipos de recursos que não fazem parte da linguagem XML, tais como anotações e dados binários, mas que estão presentes no documento e aos quais outras aplicações podem fazer referência.

XML Schema é o provável sucessor do DTD atualmente, já que é recomendado pelo W3C e possui muitas van*tag*ens sobre DTD. A primeira dessas van*tag*ens é a utilização de uma gramática mais rica e elaborada na prescrição da estrutura dos elementos, como por exemplo, especificação da quantidade exata de ocorrências possíveis de elementos filhos, de valores padrão, classificação de elementos em grupos de escolha permitindo a identificação dos elementos passíveis de serem utilizados em uma determinada localidade do documento. A segunda é van*tag*em é a formatação para digitação de informação, ou seja, é possível estabelecer máscaras para determinado valor de atributo, como, por exemplo, o número de telefone sendo composto por quatro dígitos numéricos mais o

caractere "-" mais quatro outros dígitos numéricos. A terceira e última van*tag*em do XML Schema é a existência de mecanismos de inclusão e derivação de definições que possibilitam a reutilização de definições de elementos e adaptação de definições já existentes para novas práticas.

Mas a principal diferença entre esses dois mecanismos que torna XML Schema bastante aceitável ao invés do DTD é que XML Schema utiliza a linguagem XML para codificação sintática de suas especificações. Isso simplifica o desenvolvimento de ferramentas, pois o documento e suas regras de estruturação utilizam a mesma sintaxe.

XML provém somente a sintaxe para codificação da informação do documento, sendo necessária outra ferramenta para imprimir significado a essa informação e será abordada a seguir.

Arquiteturas de Metadados

As arquiteturas de metadados visam integrar e dar suporte a uma grande variedade de esquemas de metadados espalhados em um sistema distribuído, provendo interoperabilidade sintática, semântica e estrutural da informação. Foram desenvolvidas inúmeras arquiteturas de metadados nos últimos anos e todas têm o mesmo objetivo em comum, isto é, possibilitar a troca de informações entre recursos, tais como provedores, catálogos e indexadores, e conseqüentemente prover um mecanismo de identificação e recuperação da informação mais eficiente na Web. Como contribuições importantes de arquiteturas podemos citar: Kahn e Wilensky [KAW], Warwick [LLD], *MetaContent Framework* (MCF) [GUH] e *Resource Description Framework* (RDF) [RDF]. Dentre essas, a arquitetura que mais se destaca é a RDF, elaborada pelo W3C, e que atualmente é a plataforma de desenvolvimento de aplicações na Web.

Arquitetura RDF

RDF é uma arquitetura de metadados cujo maior objetivo é definir um mecanismo de descrição de documentos que não esteja vinculado a nenhum domínio de conhecimento específico. Os mecanismos devem ter aplicação universal e ser capazes de descrever informações a respeito de qualquer tipo de domínio. Assim podem prover interoperabilidade entre aplicações através do intercâmbio de informações estruturadas de forma a possibilitar a automação de processos na Web.

A arquitetura RDF pode ser utilizada por aplicações de diversas áreas como, por exemplo [RDF]:

- Recuperação de informação: fornecendo informação estruturada de forma a possibilitar a implementação de mecanismos de pesquisa mais eficientes;
- Catalogação: descrevendo a informação e seus relacionamentos disponíveis em página web, biblioteca digital, etc; e,
- Agentes inteligentes: facilitando o compartilhamento de conhecimento e intercâmbio de informações.

RDF possui um sistema de classes, semelhante aos utilizados em sistemas de modelagem e programação orientados a objetos. Existem coleções de classes, geralmente criadas para um determinado domínio ou propósito, denominadas *schemas*. As classes são organizadas de forma hierárquica e são extensivas, ou seja, podem ser adicionadas subclasses às classes já existentes, diminuindo assim a necessidade de criação de novos esquemas. A possibilidade de compartilhamento de esquemas RDF ajuda a reutilização de definições de metadados e que juntamente com sua capacidade de extensão permite aos criados de metadados utilizar múltiplos conceitos de herança para mesclar definições, proporcionando múltiplas visões possíveis das informações e diminuindo os esforços que outros criadores teriam futuramente.

A arquitetura RDF é resultado do trabalho em conjunto de várias comunidades em torno da utilização de princípios básicos de representação e transporte de metadados na Web.

Modelo RDF

O modelo RDF [RDF] é responsável por prover um mecanismo para representação do metadado que seja neutro em termos de sintaxe e domínio de conhecimento. Ele provê a interoperabilidade estrutural, porém não fornece mecanismos para declaração e definição de propriedades e seus relacionamentos. Para a definição de propriedades de domínios específicos e sua semântica é necessária a aplicação do esquema RDF [RDFS].

O modelo RDF é utilizado para identificação de equivalência de significado, já que duas ou mais expressões em RDF são equivalentes se, e somente se, a representação de seus modelos de dados forem similares. Essa definição de equivalência permite a variação sintática em algumas expressões sem alterar seu significado. Esse modelo de dados é representado através de um DLG (*Directed Labeled Graphs*) e consiste de três tipos de objetos:

- *Resource* (Recurso): tudo que é descrito através de expressões RDF, podendo ser tanto um documento HTML, quanto um elemento XML de um documento; uma coleção de páginas ou um site inteiro. Um recurso pode também ser objeto que não seja acessado diretamente pela Web, tal como um livro impresso. Recursos são sempre nomeados por uma URI, o que permite a criação de identificadores para qualquer entidade imaginável;
- *Property* (Propriedade): é uma característica, atributo ou relação utilizado para descrever um recurso. Cada propriedade possui um significado específico, define seus próprios valores permitidos,

tipos de recursos a que podem ser aplicados e seus relacionamentos com outras propriedades; e,

- *Statement* (Declaração): é composto da associação de um recurso específico, uma propriedade e o valor da propriedade para esse recurso. Essas três partes individuais da declaração são denominadas, respectivamente, de sujeito, predicado e objeto, onde o objeto pode ser um outro recurso ou um literal, ou seja, um recurso especificado por uma URI ou uma cadeia de caracteres ou outro tipo de dados definido por XML.

Esses elementos que compõem um DLG são representados graficamente através de diagramas de nós e arcos, onde o recurso é representado por uma elipse, a propriedade por um arco ou seta e os valores por retângulos. A direção da seta é importante, já que esta sempre inicia no sujeito e aponta para o objeto da declaração.

Sintaxe RDF

O modelo RDF [RDF] fornece uma estrutura abstrata e conceitual para a definição e utilização dos metadados, mas é necessária uma sintaxe concreta para que criação e intercâmbio desses metadados seja viável. A sintaxe RDF utiliza para codificação a linguagem XML e existem dois tipos de sintaxe XML para codificar as instâncias de um modelo de dados RDF: a sintaxe de serialização, capaz de representar toda a capacidade do modelo de dados de modo simples, e a sintaxe abreviada que possui construções adicionais capazes de prover uma forma mais compacta de representação de partes do modelo de dados

Freqüentemente faz-se necessária à referência a coleções de objetos para mencionar, por exemplo, que trabalho ou material é de autoria de mais de uma pessoa ou listar os estudantes de um determinado curso. Para isso são utilizados recipientes, denominados *Containers*, que suportam uma lista de recursos ou literais. Existem três tipos de objetos RDF Container:

- *Bag*: lista não ordenada de recursos, ou literais, utilizada para declarar que uma propriedade é composta de múltiplos valores independentes da ordem de atribuição, permitindo valores duplicados. A propriedade *rdf:type* especifica o tipo de coleção que está sendo utilizado, neste caso o tipo *rdf:Bag*.
- *Sequence:* lista ordenada de recursos, ou literais, utilizada para declarar que uma propriedade pode ser composta de múltiplos valores que obedecem a uma determinada ordenação como, por exemplo, alfabética ou numérica. Este tipo de coleção também permite valores duplicados; e,
- *Alternative***:** lista de recursos ou literais que representam valores possíveis e mutuamente exclusivos para uma propriedade, proporcionando livre escolha de qualquer item da coleção.

O modelo RDF permite não apenas descrever os recursos, mas também descrever as próprias declarações (*statements*), sendo necessária uma sintaxe capaz de expressar declarações a respeito de outras declarações (*Statements about Statements*).

Esquema RDF

Na arquitetura RDF [RDF], o Modelo RDF fornece um mecanismo neutro de representação de metadados, suas propriedades e seus relacionamentos.

A codificação dessa representação é fornecida pela Sintaxe RDF, mas ainda faz-se necessário um mecanismo para definição dos recursos, suas propriedades e seus relacionamentos. Esta é a função exercida pelo Esquema RDF ou RDFS [RDFS], isto é, permitir a criação de classes de tipos de recursos e propriedades, descrições dessas classes, combinações possíveis de classes, propriedades e valores e restrições entre relacionamentos, definindo assim esquemas que podem ser utilizados em conjunto com vocabulários descritivos, tal como o Dublin Core.

Conforme já descrito, os recursos podem ser instâncias de uma ou mais classes e são indicadas pela propriedade *rdf:type*. Classes são freqüentemente organizadas de forma hierárquica. Por exemplo, uma classe denominada "Cachorro" pode ser considerada uma subclasse da classe "Mamífero" que é uma sub-classe da classe "Animal". Utilizando a notação *rdf:type*, qualquer recurso do tipo *rdf:type* Cachorro pode ser considerado também um recurso do tipo *rdf:type* Animal e assim por diante. Para especificar este tipo relacionamento entre classes é utilizada a propriedade *rdfs:subClassOf*. Além da propriedade *rdfs:subClassOf*, existem outros inúmeros recursos disponíveis para criação de declarações relativas a utilização consistente de propriedades e classes no RDF. Por exemplo, é possível que um Esquema RDF descreva as limitações de tipos de valores válidos para uma determinada propriedade, ou de propriedade válidas para uma classe. O esquema RDF é capaz de definir estes valores, mas não fornece nenhuma informação sobre quando e como um aplicativo deve processa-los.

Este sistema de tipos possíveis é especificado como recursos e propriedades no modelo de dados RDF, conforme uma hierárquica de classes, na forma de diagrama de nós e arcos do Modelo RDF. Se uma classe é um subconjunto de outra, então o arco é representado pela propriedade *rdfs:subClassOf* e sua direção têm como origem a classe principal e destino a classe secundária. Similarmente, se um recurso é uma instância de uma classe, então o arco é representado pela propriedade *rdf:type* cuja direção tem como origem o recurso e destino o nó representativo da classe

3.6. Linguagens para Criação de Ontologias

As ontologias são capazes de estabelecer uma terminologia comum entre os membros de uma determinada comunidade de interesses ou domínio de

conhecimento, sendo esses membros considerados humanos ou agentes. Elas provêm o mecanismo formal capaz de viabilizar o processamento semântico da informação através de uma máquina. Para aplicações na Web, é importante haver uma linguagem com um padrão sintático para que seja possível o intercâmbio de ontologias. Já que o XML emergiu como um padrão de linguagem para intercâmbio de informações na Web, então nada mais óbvio como o desenvolvimento de linguagens de representação de conhecimento baseadas em XML para definição de ontologias.

Diversas linguagens e mecanismos para a definição de ontologias foram criados nos últimos anos, a exemplo de: SHOE (*Simple HTML Ontology Extensions*) [SHOE], XOL (*XML-based Ontology Exchange Language*) [XOL], OIL (*Ontology Inference Layer*) [OIL], DAML (*DARP Agent Markup Language*) [DAML], dentre outros. A principal característica dessas linguagens está na capacidade de representar ontologias em RDF, arquitetura já consagrada pela W3C para interoperabilidade de informações na Web.

SHOE (Simple HTML Ontology Extensions)

A linguagem SHOE é uma extensão do HTML que permite incorporar aos documentos conteúdo com informação semântica legível pelas máquinas ou por outros documentos na Web. Recentemente, a linguagem SHOE foi adaptada para ser compatível com XML. Seu principal objetivo é possibilitar que *software*s agentes tenham acesso a informações significativas em páginas Web e documentos, melhorando os mecanismos de busca. A linguagem SHOE inclui um mecanismo de definição de ontologias, instâncias de dados em páginas Web e de classificação hierárquica de documentos HTML. Isto é feito a partir de classes e regras de restrições que especificam relacionamentos e hierarquias entre instâncias, a partir de um conjunto de *tags* acrescidos ao HTML padrão.

XOL (XML-based Ontology Exchange Language)

XOL é uma linguagem de especificação e intercâmbio de ontologias, especificado em DTD/XML. Utiliza um modelo semântico baseado em frames denominado OKBC (*Open Knowledge Base Connectivity*). Um arquivo XOL consiste de um módulo cabeçalho de definição, que provê metadados com informação sobre a ontologia, tal como nome e versão, classes e subclasses que permitem estabelecer hierarquias entre categorias de elementos, e slots que estabelecem propriedades aos elementos das classes, e definições individuais que permitem declarar nomes, descrições, informações sobre instância e valores às propriedades dos slots.

OIL (Ontology Inference Layer)

Ontology Inference Layer é uma proposta de linguagem para representação de conhecimento na Web e camadas de inferência para Ontologias que combina o uso de primitivas de modelagem de linguagens baseadas em frame com a semântica formal e serviços de dedução de proveniente de descrições lógicas. É compatível com o Esquema RDF (RDFS) [*RDF Schema*] e possui semântica precisa para descrição de significados. Uma ontologia OIL contém descrições para classes, relacionamentos, denominados slots, e instâncias. Classes podem se relacionar com outras classes através de uma hierarquia (classes/subclasses) e através de relações binárias estabelecidas entre duas relações.

Além disso, restrições de cardinalidade podem ser atribuídas aos relacionamentos. A definição de uma ontologia em OIL é constituída de dois componentes: o primeiro, denominado *ontology container*, descreve as características da ontologia, utilizando-se de descritores do padrão Dublin Core; e o segundo, denominado *ontology definitions*, define o vocabulário particular daquela ontologia.

A linguagem OIL tem sido considerada pela W3C como uma linguagem de grande relevância no contexto atual de desenvolvimento de aplicações na Web. Diante desse fato, é apresentado a seguir um exemplo, parcial, de uma ontologia definida nessa linguagem, onde parte dos termos da sintaxe é auto descritiva. Uma característica importante dessa linguagem é que a mesma pode ser utilizada em conjunto com a linguagem XML, muito embora esquemas XML não capturem totalmente a semântica embutida no OIL. Porém, sua integração com RDF Schemas (RDFS) [RDFS] é bastante promissora.

Do mesmo modo que um RDF Schema [RDFS] é utilizado para se auto definir, o mesmo também pode ser utilizado para definir outras linguagens de ontologia. Dessa forma o RDF Schema foi utilizado para definir o OIL básico, onde elementos de seu vocabulário foram mapeados para termos do Schema RDF, tais como: classe *OntologyConstraint* é mapeada como subclasse de *rdfs:ConstraintResource*, classes do tipo class-def são definidas como *rdfs:Class*, subclasses OIL *sub-classof* tornam-se *rdfs:subClassOf*, OIL-slots tornam-se sub-propriedades de rdf:Property, e assim por diante. Além disso, a sintaxe RDF inclui *namespaces* específicos para definir os termos específicos do padrão Dublin Core e outros do vocabulário OIL respectivamente, que não existam em RDF.

DAML (DARP Agent Markup Language)

A linguagem DAML é uma iniciativa da agência DARPA que está sendo desenvolvida como uma extensão de XML e RDF. A sua mais recente iniciativa é oriunda da combinação de DAML e OIL, uma linguagem que está sendo proposta como padrão para representação de ontologias e metadados pela W3C. A combinação de DAML e OIL, denominada DAML+OIL, sofre muita influência do OIL original, embora não se utilize do seu conceito original de *frames*. É constituída de uma coleção de classes e propriedades, que estão agrupados numa coleção de axiomas e precedidos pelo *tag* daml, e de objetos que

são adicionados ao RDF e RDFS. Assim, declarações (*statements*) em DAML+OIL também são declarações RDF.

A Web Semântica só poderá ser atingida a partir do inter-relacionamento automático de pequenas ontologias, desenvolvidas de forma totalmente independente e específica em seus subdomínios, servindo dessa forma como resposta a uma consulta específica. Esse fato talvez justifique o desenvolvimento e proliferação de tantas linguagens de definição de ontologias e representação de conhecimento em torno da Web Semântica. Porém, à medida que tais linguagens são utilizadas, tornam-se necessários mecanismos de edição e modelagem de ontologias que, a exemplo de ferramentas Case, possibilitem o uso de ferramentas distintas de modelagem, permitam a utilização de várias linguagens semânticas na Web.

O sistema Protégé-2000 [PROTEGE] é uma ferramenta gráfica para a edição de ontologias e aquisição de conhecimento. Inclui um mecanismo de customização que permite a modelagem conceitual em várias linguagens semânticas a exemplo de RDF, OIL e DAML+OIL. Outras ferramentas, a exemplo de OntoEdit [ONT] [SaM00] e OntoBroker [OBR] também caminham nessa direção. A OntoBroker, por exemplo, é um sistema orientado a objeto que provê compiladores em diversas linguagens para descrever ontologias, regras e fatos.

4. Conclusões

A Web Semântica pretende solucionar o problema da falta de informação semântica e significativa a respeito do conteúdo disposto na Web através de novas tecnologias, tais como, a utilização de um sistema de marcação de páginas flexível como o XML que permite incluir informações significativas a respeito de palavras ou termos do documento, uma arquitetura de metadados como RDF que padroniza a criação de metadados na Web permitindo maior intercâmbio e reutilização de descrições de termos e palavras, e Ontologias que disponibilizem um vocabulário específico de conhecimento e descrevem os termos e seus relacionamentos.

A implementação dessas novas tecnologias implica na reestruturação de páginas e web sites disponíveis atualmente na Web contemplando a criação de novas marcações e páginas de definição para termos e palavras. Surgirá, então, um novo mercado de trabalho: o de conversão de páginas web. Mas por quê as pessoas teriam custos para reestruturar seus sites Web ou páginas? O motivo desta iniciativa acredita-se que seja a criação de *softwares* agentes capazes de utilizar essas novas informações semânticas disponíveis nas páginas na execução de pesquisas mais inteligentes e eficientes e auxiliar os usuários a seguir o caminho correto para conseguir a informação desejada.

A Web Semântica não é apenas uma ferramenta para conduzir e auxiliar a execução de tarefas individuais e de pesquisas mais eficientes na Web, mas também uma ferramenta para assistir no desenvolvimento do conhecimento. Uma das maiores preocupações, hoje em dia, quando é discutido o trabalho independente de diversos pequenos grupos na Web é a necessidade de mesclar essas informações com as de outras comunidades. Um pequeno grupo é capaz de inovar de forma rápida e eficiente, mas como resultado produz uma sub-cultura cujos conceitos não podem ser compreendidos por outras pessoas. Um processo essencial para resolução desse tipo de problema é a junção de sub-culturas, conseqüente da utilização de uma linguagem em comum ou de relacionamentos e equivalências entre termos utilizados por cada grupo independente.

A Web Semântica não é uma realidade em curto prazo, mas da mesma forma que todos se sentiram surpresos com o surgimento da Web, podem se sentir também com o surgimento desta nova Web. É imprescindível que, principalmente, as empresas tenham conhecimento desta nova tecnologia e mantenham-se atentas ao início de ofertas de *softwares* agentes no mercado para que não sejam as últimas a reestruturarem seus *sites* Web.

Referências e Bibliografia Complementar

[ALTA] Altavista. http://www.av.com

[BEZ] J. Bézivin. Who's Afraid of Ontologies? http://www.metamodel.com/oopsla98-cdif-workshop/bezivin1/

[DAML] DAML: The DARPA Agent Markup Language. http://daml_about.html

[DIA] Dias, Tatiane D. *Web Semântica: Fundamentos e Tecnologias*. 2001. Trabalho de Conclusão de Curso (Graduação em Informática) - Universidade do Estado do Rio de Janeiro, 42 pg. (Unpublished).

[DC] Dublin Core Metadata Initiative. http://dublincore.org

[DOG] Dog Pile http://dogpile.com

[GRU] Grubber, T. What is an Ontology? (http://www-ksl.stanford.edu/kst/what-is-an-ontology.html)

[GUH] Guha G.V., Meta Content Framework http://mcf.research.apple.com/hs/mcf/html

[HAB] Habbib, D. P., Balliot, R. L. How to Search the World Wide Web: A Tutorial for Beginners and Non-Experts.http://204.17.98.73/midlib/tutor.htm#GSE.

[HFD] I. Horrocks, D. Fensel, J. Broekstra, S. Decker. The OntologyInference Layer
OILhttp://www.ontoknowledge.org/oil//TR/oil.long.html

[HIST] A Little History of World Wide Web. http://www.w3.org/History.html

[IAW] Iannella, R., Waugh, A. Metadata: Enabling the Internet.
http://www.dstc.edu.au/RDU/reports/CAUSE97

[INF] http://www.infoseek.com

[JHD] Hendler, J. Agents and The Semantic Web. http://www.cs.umd.edu/users/hendler/AgentWeb.html

[KAN] Kansas. Kansas City Publication Library. Introduction to Search Engines. 2001 http://www.kcpl.lib.mo.us/search/srchengines.htm

[KAW] Kahn, R. and Wilensky, R. A Framework for Distributed Object Services. http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[LLD] Lagose, C., Lynch C.; Daniel, R. The Warwick Framework – A Container Architecture for aggregating Sets of Metadata. 1996. http://www.dlib.org/ dlib/july96/lagoze/07lagoze.html

[MAG] http://www.mckinley.com

[META http://www.metacrawler.com

[OBR] http://www.ontoprise.de/start_products.htm

[OECC] Oklahoma Eletronic Commerce Connection. Semantic Web Will Force Business Site Changes. Http://www.okec.org/news/semanticweb.htm

[OIL] www.ontoknowledge.org/oil/

[ONT]http://www.ontoknowledge.org/tools/ontoedit.shtml

[PROTEGE]http://www.smi.Stanford.edu/projects/protégé/protégé-rdf/ protégé-rdf.html

[RDF] Resource Description Framework (RDF) Model and Sintax Specification. W3C Recommendation. http://www.w3.org/TR/2000/CR-rdf-schema-20000327

[RDFS] Resource Description Framework (RDF) Schemas. W3C Candidate Recommendation. http://www.w3.org/TR/rdf-schema/

[SHOE]http://www.cs.umd.edu/projects/plus/SHOE/spec.htm

[SOI] Summary Object Interchange Format (SOIF). http://harvest.cs.colorado.edu/

[TBL] Tim Berners-Lee. http://www.w3.org/People/Berners-Lee

[TBL01] Tim Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. Scientific American. http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

[UMT] University of Texas Medical Branch. Searching for Subject Information on the WWW. http://library.utmb.edu/SearchEngines/ComparisonChart.asp

[URL] Uniform Resource Location. http://www.w3.org/Addressing

[W3C] World Wide Web Consortium. http://www.w3.org

[W3C01] About World Wide Web Consortium. http://www.w3.org/Consortium

[W3C02] World Wide Web Consortium in 7 tips. http://www.w3.org/Consortium/Points

[WKL] Weibel, S. L., Kunze, J. A., Lagoze, C., Wolf, M. Dublin Core Metadata for Resource Discovery. 1998. http://www.ietf.org/rfc/rfc2413.txt

[XML] Extensible Markup Language (XML) Activity. http://www.w3.org/XML

[YAHO] Yahoo. http://www.yahoo.com.br

W3C XML Schema http://www.w3.org/XML/Schema

W3C XSL Transformation (XSLT). http://www.w3.org/TR/xslt
www.ibm.com/developer/xml

XML Schema http://www.w3c.org/XML/Schema