CARLOS ALBERTO SILVESTRE INÁCIO

WEB SEMÂNTICA E O FUTURO DA RECUPERAÇÃO *ONLINE* DA INFORMAÇÃO

Monografia apresentada à disciplina Pesquisa em Informação II como requisito parcial à conclusão do Curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná.

Orientador: Profª Denise F. Tsunoda

CURITIBA
2004

Registro aqui meus sinceros agradecimentos à professora Denise F. Tsunoda por compartilhar suas experiências profissionais e pessoais com relação à vida acadêmica, contribuindo de forma decisiva para a formação deste autor.

O conhecimento é um fenômeno complexo e multidimensional, simultaneamente elétrico, químico, fisiológico, celular, cerebral, mental, psicológico, existencial, espiritual, cultural, lingüístico, lógico, social, histórico. Oriundo necessariamente de uma atividade cognitiva, determina uma competência de ação, constituindo-se no saber que intermedia ambos processos

EDGAR MORIN

SUMÁRIO

LISTA DE FIGURAS

LISTA DE SIGLAS

CSS - Cascading Style Sheets
DTD - Document Type Definition
FTP - File Transference Protocol
HTML - Hipertext Markup Language
HTTP - Hipertext Transference Protocol
OWL - Web Ontology Language
RDF - Resource Description Framework
SGML - Standard Generalized Markup Language
SRI - Sistemas de Recuperação de Informação
SWD - Semantic Web Document
URI - Universal Resource Identifier
URL - Universal Resource Locator
W3C - World Wide Web Consortium
XHTML - Extended Hipertext Markup Language
XML - Extended Markup Language
XSL - Extensible Style Language

RESUMO

Apresenta uma introdução a mais recente tentativa de estruturação das informações disponibilizadas na internet: a *Web* Semântica. Trata de seus fundamentos conceituais e tecnológicos com enfoque principal no impacto desta estruturação na recuperação *online* da informação.

Palavras-chave: *Web* semântica; Recuperação da informação; XML (*Extensible Markup Language*); RDF (*Resource Description Framework*).

1 INTRODUÇÃO

A todo momento, pelas mais variadas razões, pessoas e empresas tomam decisões. Em todos estes casos a ação escolhida tem impacto decisivo na vida particular ou organizacional.

A intensidade deste impacto é função direta da qualidade da informação que levou à decisão. Portanto, obter a informação mais pertinente pode determinar o sucesso ou fracasso de qualquer empreendimento pessoal ou empresarial.

No que diz respeito a obtenção desta informação não se pode questionar que as novas tecnologias facilitaram ao máximo esta tarefa. O grande desafio não está na recuperação quantitativa da informação mas, e principalmente, em sua faceta qualitativa.

A Tecnologia da Informação (TI) fornece ferramentas para “fazer mais rápido” porém ainda apresenta sérias restrições quando o desejado é “fazer melhor”. E apesar da velocidade ser um fator desejável, quando ela não vem acompanhada de qualidade, serve apenas como uma comprovação mais ágil da tomada de decisão equivocada.

Com o advento e a popularização da *Internet*, mas especificamente a *web*, este problema agravou-se devido à grande descentralização da geração e disponibilização de conteúdos na grande rede.

A própria *web* procurou minimizar os problemas com a adoção de mecanismos de busca, porém estes não têm alcançado sucesso devido principalmente aos seus frágeis princípios de funcionamento centrados na recuperação de descritores, ou palavras chaves, que via de regra estão fora do contexto solicitado pelo usuário.

A principal causa deste problema é a falta de estruturação semântica dos dados disponibilizados na grande rede.

Este trabalho busca apresentar a mais recente tentativa de estruturação de conteúdos na *Internet* e seus impactos na qualidade da recuperação *online* destes conteúdos.

2 OBJETIVO

O objetivo deste trabalho é apresentar a *Web* semântica à comunidade de Gestores da Informação, através da apresentação dos seus fundamentos conceituais e tecnológicos e de suas implicações no processo de recuperação *online* da informação.

3 PROCEDIMENTOS METODOLÓGICOS

Este trabalho procurou seguir os preceitos de uma revisão de literatura com algumas particularidades ligadas ao tema: devido ao recente desenvolvimento do tema abordado, sua baixa disponibilidade de recursos em nível nacional e o suporte digital, no qual é predominantemente disponibilizado internacionalmente.

O tipo de pesquisa é: descritiva, segundo os objetivos; bibliográfica, segundo os procedimentos de coleta; documental, segundo as fontes de informação e qualitativa quanto à natureza dos dados.

Com o objetivo de sistematizar a busca por informações sobre o tema, foram adotados os seguintes procedimentos metodológicos:

a) levantamento da literatura e recursos informacionais potencialmente pertinentes;
b) leitura exploratória;
c) leitura seletiva;
d) leitura analítica;
e) tradução e condensação das informações.

3.1 FONTES DE INFORMAÇÃO SELECIONADAS

Os procedimentos metodológicos acima descritos foram aplicados sobre todos os recursos utilizados, sejam eles em suporte digital ou impresso. É importante destacar que a natureza de inovação do tema implicou na predominância do suporte digital.

Para se assegurar a qualidade das fontes em suporte digital, foram selecionadas fontes oficiais de informação sobre o tema. Por oficiais entende-se organizações legalmente constituídas e reconhecidas, com histórico de estudo científico sobre o tema , com periodicidade regular de publicações, e declarada

dedicação ao tema. Dentre estas instituições destaca-se o *World Wide Web Consortium* (*W3C*).

O *W3C* foi fundado em outubro de 1994. Instalado nas dependências do *Massachussets Institut of Technology* (MIT) em Boston. Sua missão é levar a *Web* ao seu potencial máximo, através do desenvolvimento de tecnologias (especificações, diretrizes, *software* e ferramentas) que irão criar um fórum para informação, comércio, inspiração, pensamento independente e compreensão coletiva.

Além desta fonte, foram analisadas e utilizadas quando pertinentes *sites* de instituições de ensino superior, principalmente no exterior.

No que diz respeito às instituições de ensino superior nacionais, o conteúdo disponibilizado é, em sua maioria, reapresentação de artigos internacionais.

Foram também analisados sites de instituições privadas, desta feita com maior grau de rigidez na seleção, para que se evitasse a predominância do interesse comercial sobre o científico. O mesmo princípio se adotou para a escolha dos mecanismos de busca utilizados; foram selecionados aqueles desenvolvidos e mantidos por instituições de cunho científico.

4 REVISÃO DE LITERATURA

Esta revisão de literatura se desenvolverá em duas partes. A primeira tratará dos fundamentos conceituais da *Web* semântica; a segunda abordará os seus fundamentos tecnológicos.

4.1 FUNDAMENTOS CONCEITUAIS DA *WEB* SEMÂNTICA

Para que a compreensão do fenômeno *Web* semântica possa ocorrer de maneira mais completa, se faz necessário o entendimento de conceitos fundamentais que constituem e viabilizam seu estudo.

4.1.1 Dado, Informação, Conhecimento

Dentre os conceitos mais básicos envolvidos com o tema deste trabalho, se encontram os conceitos de dado, de informação e de conhecimento. A partir deles se desenvolvem todos os estudos que levam ao entendimento da *Web* semântica.

Para SETZER (2001) “dado é uma seqüência de símbolos quantificados ou quantificáveis. Desta forma, o texto é uma espécie de dado. Na verdade, caracteres são quantificados porque existem em número finito. Imagens, sons gravados e animações também são considerados dados porque podem ser quantificáveis”.

KASABOV (1996, p.75) descreve dado como “matéria bruta, sem significado contextual”. Quando este mesmo dado é contextualizado, organizado em grupos e estruturas, são chamados de informação. Desta forma, segundo KASABOV, informação pode ser definida como qualquer dado estruturado que apresente significado contextual.

No que tange ao conceito de informação, SETZER (2001) diz: “informação é uma abstração informal”. Em outras palavras, “a informação não pode ser formalizada através da lógica ou da matemática. A causa principal para essa impossibilidade é a presença da subjetividade”. O significado da informação depende da interpretação pessoal. Quando se representa a informação através de dados, esta representação poderá ser armazenada. Porém é importante destacar que o que foi armazenado não foi a informação em si, mas sua representação na forma de dados.

Para BIOLCHINI (2001, p.1) o fenômeno da informação encontra-se estreitamente vinculado aos processos da cognição humana. E acrescenta (...) a informação tem sido considerada como elemento fundamental para o conhecimento, condição mesmo de sua possibilidade.

Para uma definição de conhecimento encontra-se em KASABOV: “é informação em alto grau de estruturação. Ele é informação condensada. É a representação concisa de uma experiência.” KASABOV argumenta ainda que o conhecimento está expresso nas regras que utilizamos para agir durante nosso cotidiano.

Avançando um pouco pelo aspecto filosófico da questão, encontramos ABBAGNANO (1970), para o qual conhecimento é “como um procedimento operacional, uma técnica de verificação de um objeto qualquer, isto é, qualquer procedimento que torne possível a descrição, o cálculo ou a previsão controlável de um objeto; e por objeto deve entender-se qualquer entidade, fato, coisa, realidade ou propriedade, que possa ser submetido a um tal procedimento”.

Para MORIN (1986), o conhecimento “é um fenômeno complexo e multidimensional, simultaneamente elétrico, químico, fisiológico, celular, cerebral, mental, psicológico, existencial, espiritual, cultural, lingüístico, lógico, social, histórico. Oriundo necessariamente de uma atividade cognitiva, determina uma competência de ação, constituindo-se no saber que intermedia ambos processos”.

Corroborando o conceito acima, BIOLCHINI (2001, p.8) complementa que “ ... a presença de fenômenos informacionais se dá de forma constitutiva em diversos processos da cognição, tais como aqueles relacionados à percepção, à compreensão, à memória, à resolução de problemas, à elaboração, execução e monitoramento de ações, à comunicação entre humanos e máquinas ou sistemas artificiais, e à aquisição de novos conhecimentos e habilidades”.

4.1.2 Representação do Conhecimento

Para resolver os problemas complexos – leia-se, tomar decisões com relação a eles - via de regra é necessário uma grande quantidade de conhecimentos e principalmente o emprego de estratégias e mecanismos para manipulá-los.

Quando se discute representação do conhecimento trata-se com dois tipos diferentes de entidades: os fatos (o que queremos representar); e a representação dos fatos (o que seremos capazes de manipular).

4.1.2.1 Conceito de representação do conhecimento

Representação do conhecimento pode ser definido como um conjunto de convenções sintáticas e semânticas que torna possível descrever coisas.

O conhecimento poderá ser representado de diferente maneiras em função de suas características e dos mecanismos aptos para representa-lo. A escolha do mecanismo ou técnica mais adequada é determinante para que se chegue o mais próximo possível do fato que se pretende analisar.

O dado facilita o processamento automático, enquanto que o conhecimento exige maiores esforços para o seu processamento.

Este esforço normalmente está ligado à estruturação dos dados. A Ciência da Informação, que lida com o tratamento e recuperação da informação, utiliza preferencialmente o recurso do metadado para a estruturação da informação.

4.1.3 Metadados

Metadados são de fundamental importância na Ciência da Informação pois se reportam diretamente à maneira como os dados estão estruturados e à sua forma de recuperação.

De uma maneira geral, os autores conceituam metadados como “dados sobre dados”. Adotando um pouco mais de rigor, TAKAHASHI (2000) atribui a metadados o conceito “são dados a respeito de outros dados, ou seja, qualquer dado usado para auxiliar na identificação, descrição e localização de informações. Trata-se, em outras palavras de dados estruturados que descrevem as características de um recurso de informação".

Considerando uma visão mais pragmática MILSTEAD e FELDMAN (2004) definem metadados como dados associados com objetos que aliviam seus usuários potenciais da tarefa compreender a complexidade de sua estruturação e características, se concentrando em sua utilização.

Os avanços em tecnologia da informação têm provocado mudanças no comportamento dos profissionais da informação. OLIVEIRA (2002) destaca “o domínio da representação descritiva e de conteúdos (...) recebe impacto, não apenas pela multiplicidade ou variedade de suportes informacionais mas pela inserção de novos atores e novas tecnologias no ambiente informacional”.

4.1.3.1 Aplicações de metadados

São inúmeras as aplicações de metadados; na catalogação, na descoberta de recursos, no comércio eletrônico, nos aplicativos com agentes inteligentes, assinaturas digitais, no direito autoral e na avaliação de conteúdo, na preservação digital, etc.

Apesar de existirem argumentações contra a utilização de metadados no meio digital – devido à característica digital dos conteúdos – LAGOZE (2001) argumenta em defesa de sua utilização:

a) o tamanho dos substitutos em geral é menor do que o objeto real, sendo mais prático a manipulação do substituto nos processos de descoberta;
b) por questões de propriedade intelectual, pode ser mais interessante para os detentores do direito de propriedade o acesso aos substitutos do que ao recursos real.

Ainda segundo LAGOZE (2001), metadados podem ser pensados da forma como auxiliam a construir múltiplas visões sobre um único objeto de informação. Estas visões podem compor, por exemplo, a base para serviços de informações.

Para WEIBEL (1995), a estruturação de metadados deve seguir princípios, a saber:

a) intrinsicalidade: o metadado deve se concentrar em descrever propriedades intrínsecas do objeto;

b) extensibilidade: possibilita ao usuário inserir material descritivo para atender a uma necessidade específica;
c) sintaxe independente: isto facilita o uso do metadado em disciplinas e aplicações diferentes;
d) opcionalidade: todos os elementos devem ser opcionais;
e) repetibilidade: todos os elementos podem ser repetidos;
f) modificabilidade: um elemento pode ser alterado para atender a uma comunidade em especifico.

4.1.4 Hipertexto

Este termo foi cunhado pela primeira vez por volta de 1965. Originalmente o termo serviu para designar "escrita não-seqüencial" depois evoluiu para além do texto, incluindo outros recursos tais como imagens e sons.

Atualmente hipertexto identifica palavras ou expressões que remetem a outra posição no próprio texto ou para outro documento numa rede.

Para SMITH (2001), hipertexto é uma abordagem da gestão de informação na qual os dados são armazenados em uma rede de nós conectados por ligações. Os nós podem conter textos, gráficos, áudio e vídeo, bem como programas de computador ou outras formas de dados.

A *Web* é essencialmente hipertexto. Ele possibilita a não-linearidade, a liberdade para o usuário definir seu próprio caminho.

Segundo LANDOW (1992), o hipertexto põe em cheque: seqüências fixadas, começo e fim definidos. Na narrativa hipertextual, o autor oferece múltiplas possibilidades através das quais os próprios leitores constróem sucessões temporais e escolhem personagens, realizando saltos com base em informações referenciais.

SCHNEIDERMAN e KEARSLEY (1989), definem hipertexto como "uma rede de nós e ligações entre documentos, onde os documentos são nós e as ligações são referências cruzadas." Ainda segundo os autores as redes podem apresentar uma hierarquia e os nós não se restringem a textos, mas podem ser gráficos, fotos, sons narração ou seqüência animadas (vídeo).

Pode-se considerar um conceito resultante da extensão do termo hipertexto, quando os documentos são de natureza (tipo) multimeios. Neste caso admite-se o termo hipermídia.

Com o avanço da qualidade e capacidade de *softwares* e *hardwares*, a *Internet* apresenta-se cada vez mais hipermídia.

4.1.5 Linguagens de Marcação de Texto

Além da utilização de metadados e hipertexto, um documento bem estruturado requer principalmente a utilização de linguagens de marcação de texto, que são metalinguagens capazes de demarcar estruturas e conteúdos de um recurso informacional disponível em suporte digital.

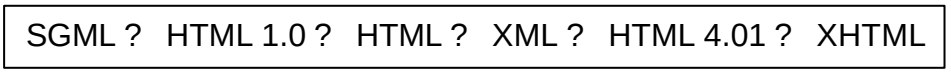
A linguagem mais conhecida na atualidade é a HTML (*Hypertext Markup Language*), voltada para estruturação de documentos e para a apresentação visual de documentos em um navegador.

Porém, HTML pertence a uma família extensa e não é mais considerada a mais importante dentre elas, principalmente quando o que se pretende é a marcação de conteúdo em detrimento da marcação simples de texto.

HTML é derivada da linguagem pioneira de marcação SGML (*Standard Generalized Markup Language*) e foi criada por Tim Berners Lee - idealizador da *Web* - especificamente para a composição e apresentação de documentos na *Web*.

A grande família da qual a HTML faz parte é a seguinte:

FIGURA 1 – A FAMÍLIA DO HTML



FONTE: FARIA e GIRARDI (2004)

Na atualidade, em função da crescente demanda pela estruturação de documentos com vistas a explicitar seu conteúdo e não apenas suas formas, a utilização da XML (*Extensilble Markup Language*) tem se consolidado como mais importante.

A utilização do hipertexto em conjunto com as linguagens de marcação de texto favoreceram o surgimento de sistemas de recuperação da informação mais eficazes.

4.1.6 Sistema de Recuperação de Informação (SRI)

Para se chegar ao conceito de SRI é importante buscar, em primeiro lugar a definição de sistema num contexto mais amplo. Isto facilitará sua compreensão e aplicação ao campo da recuperação da informação.

BERTALANFFY (1976) foi o precursor da teoria dos sistemas. Sua intenção inicial era definir uma metodologia para o tratamento dos problemas científicos. Para ele “sistema é um conjunto organizado de partes inter-atuantes e interdependentes, que se relacionam formando um todo unitário e complexo”.

O autor entende a ciência como um “subsistema do sistema conceitual, definido-a como um sistema abstrato, isso é, um sistema conceitual correspondente à realidade”.

A meta final da Teoria Geral dos Sistemas não é buscar analogias entre as ciências, mas evitar a superficialidade científica.

As sucessivas especializações das ciências levam à criação de jargões que fazem sentido apenas para uma comunidade científica em especial. Este fato tem gerado dificuldades principalmente no desenvolvimento de projetos interdisciplinares. A Teoria Geral dos Sistemas visa introduzir uma semântica científica de utilização universal.

É importante esclarecer que as partes que compõem um sistema não se referem ao campo físico (objetos), mas sim ao campo funcional. Deste modo estas partes passam a ter funções básicas realizadas pelo sistema.

Podemos enumerar as partes de um sistema como entrada, processamento, caixa preta, saída, retro-alimentação.

A Entrada se refere à alimentação do sistema. Em seguida tem-se o Processamento desta Entrada. A caixa preta identifica algum processamento – não explícito - sofrido pela Entrada. Segue-se então até a Saída que apresenta o resultado do Processamento. Finalmente ocorre a retro-alimentação, que leva os dados processados obtidos na saída para a Entrada.

4.1.6.1 Conceito de SRI

Segundo LANCASTER e WARNER (1993, p.4-5), Sistemas de Recuperação de Informação “são uma *interface* entre uma coleção de recursos de informação - em meio impresso ou não - e uma população de usuários. Desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle destes; distribuição e disseminação aos usuários”.

As características conceituais atribuídas aos SRIs em geral, aplicam-se também aos Sistemas de Recuperação *online* de Informação. Atentando-se para o detalhe da mudança do ambiente. No caso dos SRIs *online*, restringe-se este ambiente digital, normalmente com utilização de redes de computadores, principalmente a *world wide web*.

4.1.7 A *World Wide Web* (www)

A *World Wide Web* (também chamada *web* ou www) é a interface gráfica da *Internet*. Ela é um sistema de informações que visa englobar todos os outros sistemas de informação disponíveis na *Internet*.

Sua idéia básica é criar um mundo de informações sem fronteiras, prevendo as seguintes características: *interface* consistente; incorporação de um vasto conjunto de tecnologias e tipos de documentos; "leitura universal".

Para isso, utiliza três importantes tecnologias: um protocolo de transmissão de dados HTTP; um sistema de endereçamento próprio URL; uma linguagem de marcação, para transmitir documentos formatados através da rede HTML ou XML.

HTTP significa *HyperText Transfer Protocol* - Protocolo de Transferência de Hipertexto. O HTTP é o protocolo usado para a transmissão de dados no sistema *World Wide Web*. Cada vez que você aciona um *link*, seu navegador (*browser)* realiza uma comunicação com um servidor da *Web* através deste protocolo.

O sistema de endereçamento da Web é baseado em uma sintaxe chamada URI (*Universal Resource Identifier* - Identificador Universal de Recursos). Dentre os vários tipos de URI’s, destacam-se atualmente os URLs (*Uniform Resource Locator*) largamente empregados na *Web* para identificar recursos a partir de sua localização na rede. Em exemplo de sua sintaxe:

http://www.ufpr.br/decigi/curso/ementas.html

Esse endereço identifica:

a) o protocolo de acesso ao recurso desejado: http;
b) a máquina a ser conectada e o local: www.ufpr.br;
c) o caminho de diretórios até o recurso: /decigi/curso/;
d) o recurso (arquivo) a ser obtido: ementas.html.

4.1.8 Motores de busca (*Search Engines*)

Devido a característica altamente descentralizada da *web*, os conteúdos foram disponibilizados de forma não estruturada, sem uma preocupação com a maneira de recupera-los quando necessário.

Como tentativa de sistematizar estas buscas foram desenvolvidos diversos *Search Engines* (motores de busca). Encontram-se disponíveis hoje na *Web*, motores de busca especializados por temas, público alvo etc. Estes motores percorrem a *Web* em busca da informação solicitada normalmente por meio de descritores (palavra ou conjunto de palavras que identificam o que se quer recuperar).

A despeito da grande quantidade e variedade de motores de busca, a sistemática de indexação e apresentação dos resultados não sofre alteração significativa entre um motor e outro. De uma maneira geral estes motores varrem a rede a partir do recebimento de um descritor.

Alguns buscam prioritariamente títulos de documentos, outros além disso, analisam a quantidade de vezes que o documento foi buscado/citado em um determinado espaço de tempo e contexto específico.

Porém, invariavelmente, todos partem de um descritor e buscam a ocorrência deste na *web*. Técnica que tem se mostrado ineficaz por retornar uma quantidade extremamente grande de respostas, muitas delas sem nenhuma relação com a solicitação do usuário. Na medida em que cresce vertiginosamente o conteúdo disponível, aumentam proporcionalmente a quantidade de respostas equivocadas.

O fator preponderante na determinação destes resultados equivocados está na incapacidade que os aplicativos apresentam de “entender” semanticamente o

descritor solicitado. Os motores de busca carecem de semântica. Apenas com a capacidade de entender semanticamente os descritores, os motores poderão chegar a resultados considerados desejáveis para cada solicitação de busca.

A grande promessa de solução deste problema está na *Web* semântica.

4.1.9 A Evolução da *Web*

A evolução da *Web*, segundo JASPER e USCHOLD (2001), pode ser caracterizada segundo várias perspectivas:

a) localização de recursos: a forma como as pessoas encontram recursos na *Web* está evoluindo do simples texto e palavras-chaves para uma busca mais sofisticada, envolvendo técnicas semânticas para buscas e navegação;
b) usuários: os recursos disponíveis estão evoluindo de maneira a serem projetados para consumo humano e de máquinas simultaneamente;
c) serviços e funções da Web: a Web está evoluindo de um lugar para se encontrar recursos para um lugar para se criar recursos.

4.1.10 Semântica

Semântica é o estudo da relação de significação nos signos e da representação do sentido dos enunciados. FERREIRA (1999)

Alguns autores da área de ciência da computação, como TRIPPE (2001) , consideram que “ algo tem semântica quando pode ser processado e entendido por um computador “.

A apropriação do termo pela Ciência da Computação implica em nova significação para atender às suas necessidades conceituais.

4.1.11 Ontologias

Ontologia é um termo da filosofia que significa “o estudo do ser enquanto ser”. Adaptado para a Ciência da Computação, passou a designar “uma coleção de informações com uma especificação formal e consensual de conceitos, definindo as relações entre os conceitos e que providencia um compartilhamento e um entendimento comum de um domínio que pode ser comunicado entre pessoas e sistemas de aplicações”. (HENDLER, 2003)

4.2 FUNDAMENTOS TECNOLÓGICOS DA WEB SEMÂNTICA

Adotou-se por fundamentos tecnológicos aqueles que dinamizam a aplicação do conceito, determinando sua aplicabilidade.

Os pilares que fundamentam o desenvolvimento da *Web* Semântica na atualidade são a metalinguagem XML (*eXtended Markup Language*), o RDF (*Resource Description Framework*) e o desenvolvimento de Ontologias.

4.2.1 XML

Extensible Markup Language (XML) é uma linguagem desenvolvida para a descrição de dados (conteúdo). Ela permite a criação de formatos únicos para a descrição de dados de aplicações específicas.

A linguagem XML possui a importante característica de ser extensível, permitindo que novas *tags* (elementos) de marcação sejam criadas por quem a utiliza.

XML é destinada à descrever o conteúdo de um documento, e a linguagem HTML tem como objetivo definir a formatação do mesmo, ou seja, o XML define o contéudo, e o HTML define como ele será exibido ao usuário.

4.2.1.1 Os principais benefícios do XML

O XML permite múltiplas formas de visualização. Isso permite que um único documento possa ser apresentado de diversas formas, de acordo com o gosto do usuário ou de acordo com as configurações da aplicação em uso. Essa múltipla visualização é processada localmente, no cliente.

O XML permite a integração de dados estruturados de diversas fontes, tais como bancos de dados. Essa integração pode ser feita em um servidor intermediário, e os dados estarão disponíveis para clientes ou outros servidores.

Por ser extensível, o XML pode descrever dados de uma enorme variedade de aplicações (registro de dados, notícias, transações comerciais, etc...) e por possuir *tags* autodescritivas não precisa de uma descrição de contexto acoplada ao documento como o HTML.

AMARAL (2003), traduzindo um documento oficial do W3C em uma linguagem direta e coloquial, apresenta - em dez pontos – as características principais da linguagem XML:

1. XML é para estruturar dados.

   São exemplos de dados estruturados planilhas, cadernos de endereços, parâmetros de configuração, transações financeiras e desenhos técnicos. XML é um conjunto de regras (você também pode encará-las como convenções ou diretrizes) para projetar formatos de texto que o permitam estruturar seus dados. XML não é uma linguagem de programação e você não precisa ser um programador para usá-la ou aprendê-la. XML torna simples para o computador gerar e ler dados, e garantir que sua estrutura não seja ambígua. XML evita os problemas mais comuns em projetos de linguagens; ela é extensível, independente de plataforma e suporta internacionalização e localização.

2. XML parece um pouco com HTML.

   Como HTML, XML usa marcadores (palavras envoltas pelos sinais '<' e '>') e atributos (na forma nome="valor"). Mas enquanto HTML especifica o que cada marcador e atributo significa, e às vezes como seu conteúdo aparecerá num navegador, XML usa os marcadores apenas para delimitar os trechos de dados, deixando sua interpretação completamente à cargo da aplicação que os lê. Em outras palavras, ao ver "<p>" num arquivo XML, não assuma que é um parágrafo. Dependendo do contexto, pode ser um preço, um parâmetro, uma pessoa, um p... — ora, quem disse que a palavra tem que começar por “p”?

3. XML é texto, mas não é para se ler.

   Programas que produzem planilhas, listas de endereços e outros dados estruturados freqüentemente os gravam em disco, usando um formato binário ou textual. Uma vantagem do formato textual é que ele permite às pessoas, se necessário, ver os dados sem usar o programa que os produziu; ou seja, você pode ler um formato textual com o seu editor de textos favorito. Formatos textuais também ajudam os desenvolvedores a depurar mais facilmente as aplicações. Como em HTML, os arquivos XML são arquivos-texto que as pessoas não deveriam precisar ler, mas podem fazê-lo em caso de necessidade. A semelhança diminui quando vemos que as regras para arquivos XML são rígidas. Um marcador esquecido ou um atributo sem aspas inutilizam um arquivo XML, enquanto em HTML tal prática é tolerada e com freqüência explicitamente permitida. A especificação oficial da XML proíbe as aplicações de tentar inferir a intenção do autor de um arquivo defeituoso; se um defeito é encontrado, a aplicação é obrigada a parar ali mesmo e sinalizar um erro.

4. XML é prolixo de propósito.

   Como XML é um formato textual e usa marcadores para delimitar os dados, os arquivos XML são quase sempre maiores que num formato binário equivalente. Isso é fruto de uma decisão consciente dos projetistas da XML. As vantagens de um formato textual são evidentes (vide item 3), e as desvantagens podem ser geralmente compensadas num outro nível. Espaço em disco já não é tão caro como costumava ser, e programas de compressão como zip e gzip podem comprimir arquivos rápida e eficientemente. Além disso, protocolos de comunicação modernos e o HTTP/1.1, o protocolo central da Web, podem comprimir os dados em trânsito, poupando banda tão eficientemente quanto um formato binário.

5. XML é uma família de tecnologias.

   XML 1.0 é a especificação que define o que são "marcadores" e "atributos". Além de XML 1.0, a "família XML" é um conjunto crescente de módulos que oferecem serviços úteis para levar a cabo tarefas importantes e muito requisitadas. Xlink descreve uma forma padronizada de inserir hiperlinques num arquivo XML. XPointer e XFragments são sintaxes em desenvolvimento para endereçar partes de um documento XML. Um XPointer parece com um URL, mas ao invés de apontar para documentos na Web, ele aponta para trechos de dados dentro de um arquivo XML. CSS , a linguagem de folhas de estilo, aplica-se tanto a XML como a HTML. XSL é uma linguagem avançada para expressar folhas de estilo. Ela é baseada em XSLT, uma linguagem de transformação usada para rearranjar, adicionar ou apagar marcadores e atributos. O DOM é um conjunto padrão de funções para manipular arquivos XML (e HTML) com uma linguagem de programação. As recomendações Esquema XML 1 e 2 ajudam os desenvolvedores a definir precisamente as estruturas de seus próprios formatos baseados em XML.

6. XML é novidade, mas nem tanto assim.

   O desenvolvimento de XML começou em 1996 e é uma Recomendação W3C desde fevereiro de 1998, o que pode levá-lo a crer que XML é uma tecnologia imatura. Na verdade, esta tecnologia não é muito recente. Antes de XML já existia SGML, desenvolvida no início da década de 80 e padrão ISO desde 1986, largamente utilizada em grandes projetos de documentação. O desenvolvimento de HTML começou em 1990. Os projetistas da XML simplesmente pegaram as melhores partes da SGML, guiados pela experiência acumulada com HTML, e produziram algo que não é em nada menos poderoso que SGML, e amplamente mais regular e simples de usar. Contudo, às vezes é difícil distinguir algumas evoluções de revoluções... E deve ser dito que, enquanto SGML é usada principalmente para documentação técnica e muito menos para outros tipos de dados, com XML ocorre exatamente o oposto.

7. XML leva a HTML à XHTML.

   Há uma importante aplicação XML que é um formato de documento: é a XHTML, a sucessora da HTML. XHTML tem muitos dos mesmos elementos que HTML, mas a sintaxe foi ligeiramente modificada para se conformar às regras da XML. Uma aplicação que é "baseada em XML" herda a sintaxe de XML e a restringe de certas formas (e.g., XHTML aceita "<p>", mas não "<r>"); ela também acrescenta significado à sintaxe (XHTML reza que "<p>" significa "parágrafo", e não "preço", "pessoa" ou qualquer outra coisa).

8. XML é modular.

   XML permite que você defina um novo formato de documento combinando ou reutilizando outros formatos. Como dois formatos desenvolvidos independentemente podem ter elementos ou atributos homônimos, deve-se ter cuidado ao combinar tais formatos ("<p>" significa "parágrafo" deste formato ou "pessoa" daquele outro?). Para eliminar a confusão de nomes ao combinar formatos, XML provê um mecanismo de espaços nominais (namespaces). XSL e RDF são bons exemplos de formatos que usam espaços nominais. O Esquema XML foi projetado para reproduzir este suporte à modularidade no nível da definição da estrutura dos documentos XML, tornando fácil combinar dois esquemas para produzir um terceiro que represente uma estrutura de documento híbrida.

9. XML é a base de RDF e da *Web* Semântica.

   A Framework para Descrição de Recursos (RDF) é um formato textual XML para descrever recursos e aplicações de metadados, como listas de reprodução de músicas, álbuns de fotos e bibliografias. Por exemplo, RDF permite que você identifique pessoas num álbum de fotos na Web usando informação de uma lista de contatos pessoais; assim, o seu programa de correio poderia automaticamente disparar uma mensagem para essas pessoas dizendo que suas fotos estão disponíveis na Web. Assim como HTML integrou documentos, sistemas de menu e formulários para deslanchar a Web original, RDF integra aplicações e agentes numa Rede (Web) Semântica. Assim como as pessoas precisam concordar acerca do significado das palavras que utilizam para se comunicar, os computadores também necessitam pactuar o significado dos termos para poderem se comunicar efetivamente. A descrição formal dos termos de uma certa área (comércio ou manufatura, por exemplo) são denominadas ontologias e são uma parte vital da Web Semântica. RDF, ontologias e a representação formal do significado, de modo que os computadores possam ajudar as pessoas em seus trabalhos, são todos tópicos em discussão no grupo Semantic Web Activity.

10. XML é livre de licenças, independente de plataforma e bem suportada.

    Ao basear um projeto em XML, você herda um vasto e crescente conjunto de ferramentas (uma das quais pode fazer exatamente o que você precisa!)

> e uma comunidade de engenheiros com experiência na tecnologia. Optar por XML é semelhante a escolher SQL para bancos de dados: você ainda tem que montar sua própria base de dados e os programas e rotinas para manipulá-la, mas há muitas ferramentas disponíveis e muita gente que pode ajudá-lo. E como XML é livre de licenças, você pode criar seu próprio software com ela sem ter que pagar nada a ninguém por isso. O vasto e crescente suporte significa que você não está preso a um simples fornecedor. XML não é sempre a melhor solução, mas vale sempre a pena considerá-la.

4.2.2 *Resource Description Framework* (RDF)

O RDF (*Resource Description Framework*) é uma metalinguagem desenvolvida com o objetivo de definição e utilização de metadados. O RDF, através de sua sintaxe, utiliza o XML para expressar o significado da informação. Desta maneira RDF e XML trabalham em conjunto para melhor identificar um conteúdo.

Este conteúdo é representado através de um conjunto de três informações denominado Tripla. As informações que compõem a tripla designam um *objeto*, um *atributo* e um *valor*. A formatação da tripla obedece a seguinte ordem A(O,V), isto é, um objeto "O" tem um atributo "A" cujo valor é "V".

FARIA e GIRARDI (2004) apresentam exemplos de algumas triplas:

As Figuras abaixo mostram exemplos de três relacionamentos no formato das triplas de A(O,V):

FIGURA 2 – TRIPLA "A"

hasName ('http://www.w3.org/employee/id1321', "Jim Lerners")

FONTE: FARIA e GIRARDI (2004)

Significa:

O objeto	*http://www.w3.org/employee/id1321*
tem um atributo	*hasName*,
cujo valor é	*Jim Lerners*

> Em outras palavras, o objeto em questão é um URL que identifica uma instituição, no diretório de empregados, indicando o empregado cuja Id é 1321, cujo nome é Jim Lerners.

FIGURA 3 – TRIPLA “B”

authorOf (‘http://www.w3.org/employee/id1321’,‘http://www.books.org/ISBN12515866’)

FONTE: FARIA e GIRARDI (2004)

Significa:

O objeto	*http://www.w3.org/employee/id1321*
tem um atributo	*authorOf*
cujo valor é	*http://www.books.org/ISBN12515866*

Em outras palavras, o objeto é um URL que identifica um empregado que é autor de um livro cujo ISBN é 12515866, e está disponível em outro URL

FIGURA 4 – TRIPLA “C”

hasPrice (‘http://www.books.org/ISBN12515866’, “$62”).

FONTE: FARIA e GIRARDI (2004)

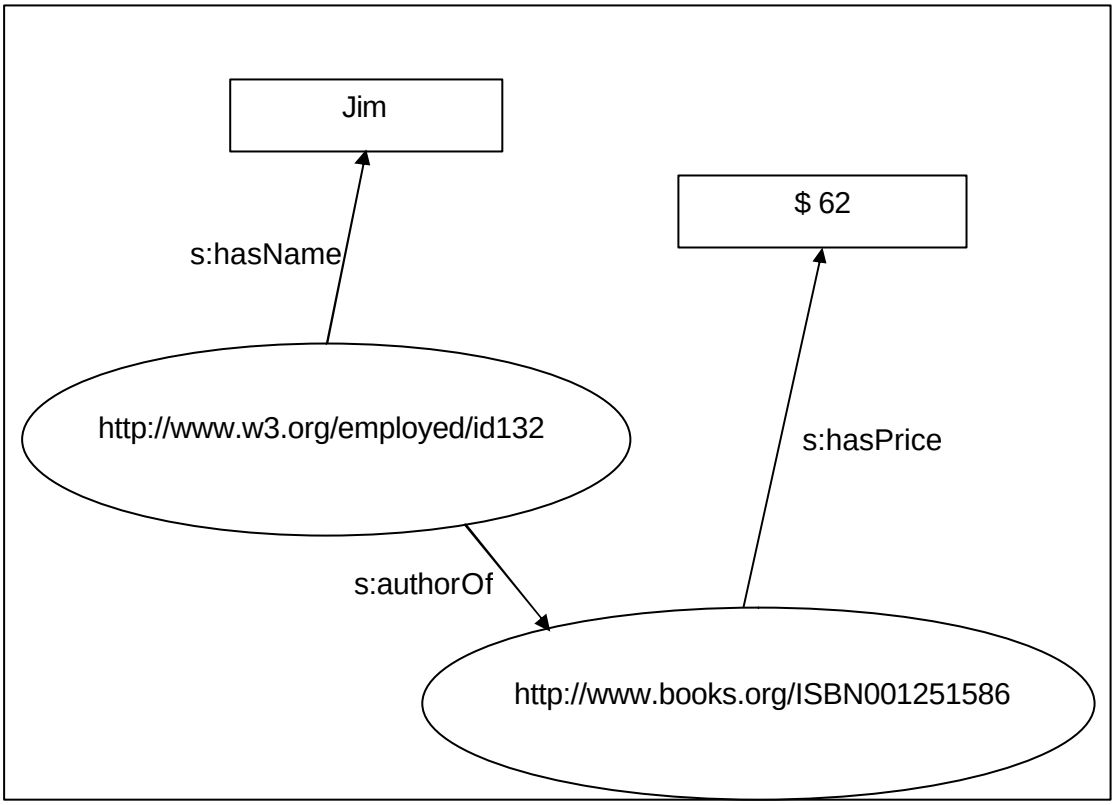
Significa:

O objeto	*http://www.books.org/ ISBN12515866*
tem um atributo	*hasPrice*
cujo valor é	*“$62.*

Em outras palavras, o objeto é um URL que identifica um livro cujo ISBN é 12515866 e seu preço é *“$62*.

O formato A(O,V) é empregado para o processamento. Porém, há uma outra possibilidade que - apesar de não se prestar para a computação - oferece uma melhor visualização do conteúdo representado. Esta maneira de representar recebe a denominação de Grafo. A representação da “Tripla C” acima em forma de grafo teria a configuração apresentado abaixo.

FIGURA 5 – EXEMPLO DE UM GRAFO RDF



FONTE: FARIA e GIRARDI (2004)

4.2.3 Ontologia

Uma ontologia é uma especificação de uma conceituação. É designada com o propósito de habilitar o compartilhamento e reuso de conhecimento, de forma a criar “compromissos ontológicos”, ou definições necessárias à criação de um vocabulário comum. SOUZA (2004).

Segundo FARIA e GIRARDI (2004), para a construção de uma ontologia são utilizados os seguintes objetos:

a) entidades: descrevem conceitos (elementos de um domínio estudado) e providenciam uma representação lógica;
b) atributos: descrevem as propriedades das entidades;
c) relações: descrevem as ligações entre os objetos no modelo (entidades e atributos);
d) restrições: são condições que o projetista impõe sobre as entidades, atributos ou relações;

Ainda segundo os mesmo autores, uma ontologia possui uma hierarquia de conceitos dentro de um domínio, as descrições de cada conceito e as propriedades definidas por atributos de tipo valor. Geralmente consiste de uma taxonomia e de um conjunto de regras de inferências. Uma taxonomia define classes, subclasses e as relações entre elas.

O W3C recomenda o uso da *Web Ontology Language* (OWL) para definição de ontologias, que permite definições de *slots* e definições de classes. A definição de *slot* descreve um relacionamento binário entre duas entidades. A definição de classe associa o nome da classe com a definição da classe.

Para FARIA e GIRARDI (2004), as definições de classes têm os seguintes componentes:

a) o tipo de definição pode ser do tipo “definida” (em que na sua definição ela é completamente específica), ou pode ser do tipo “primitiva” (em que a definição da classe é necessária, mas insuficiente para determinar a associação da classe);
b) o slot restrição (slot constraint) restringe o valor que o slot pode assumir, quando for aplicado a uma instância de uma classe;
c) as subclasses (subclass of) relacionam uma classe definida previamente com uma lista de uma ou mais classes.

4.2.3.1 Exemplo de ontologia

As ontologias são determinantes no ambiente *Web* semântica, pois cabe a elas a formalização de um conceito de uma determinada área do conhecimento para uma posterior utilização por máquinas e humanos.

Abaixo encontra-se uma amostra de uma ontologia que formaliza a vida selvagem africana. Esta formalização preocupa-se em deixar o mais claro possível, sem ambigüidades, o conceito em questão através da utilização de classes próprias de cada linguagem, por exemplo a OWL.

FIGURA 6 – ONTOLOGIA DEFININDO A VIDA SELVAGEM AFRICANA

```
class-def animal
class-def plant
subclass-of NOT animal
class-def tree
subclass-of plant
class-def branch
slot-constraint is-part-of
has value tree
class-def leaf
slot-constraint is-part-of
has value branch
class-def defined carnivore
subclass-of animal
slot-constraint eats
value type animal
class-def defined herbivore
subclass-of animal
slot-constraint eats
value type plant
OR (slot-constraint is-part-of
has value plant)
class-def giraffe
subclass-of herbivore
slot-constraint eats
value type leaf
class-def lion
subclass-of animal
slot-constraint eats
value type herbivore
class-def tasty-plant
subclass-of plant
slot-constraint eaten-by
has value herbivore, carnivore
```

FONTE: FARIA e GIRARDI (2004)

Como pode ser verificado na figura acima, a ontologia expressa regras passíveis de serem manipuladas por uma máquina.

FARIA e GIRARDI (2004) argumentam que uma página XML pode ligar-se a uma ontologia sob forma de DTD (*Document Type Definition)*, onde são definidos os significados de cada uma das *tags* utilizadas. Com páginas da Web ligadas a uma ontologia se tornará mais simples a resolução de problemas de indefinição ou conflito de terminologia, melhorando o armazenamento, compartilhamento e

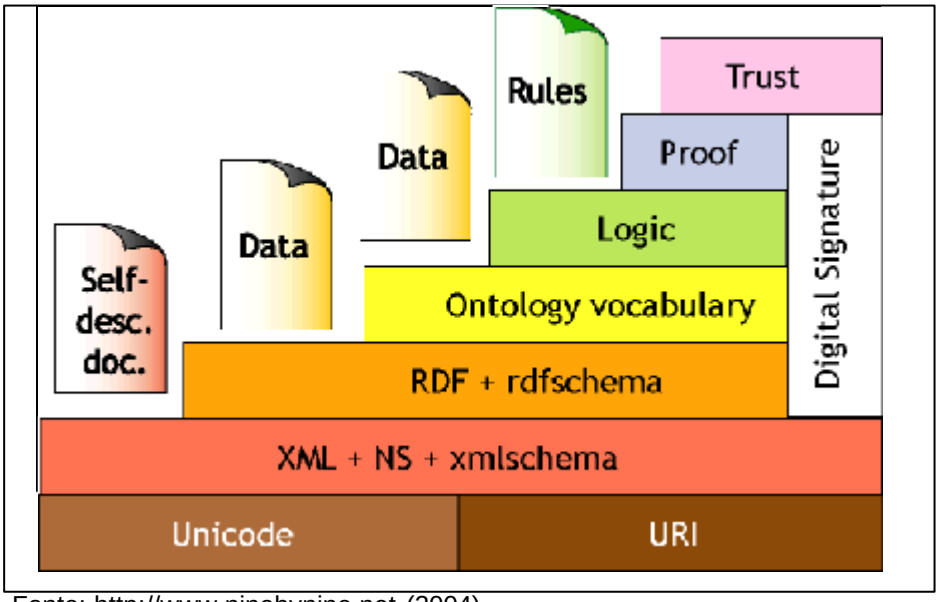
processamento do conhecimento. Uma página pode ser relacionada com outras utilizando as regras de inferência definidas na camada lógica.

Afim de demonstrar a utilidade das ontologias na resolução de ambigüidades ou autores acrescentam: “as ontologias possuem relações de equivalência, que resolvem o problema de existirem palavras sintaticamente diferentes, mas similares semanticamente, ou seja, palavras escritas diferentes, mas que possuem significados similares. Um exemplo seria um site que tem disponibilizado um código postal do seu criador, e outro site que tem um CEP disponibilizado, através do uso das ontologias se tornará claro que CEP e código postal são palavras que contém o mesmo significado apesar de estarem escritos de formas diferentes”.

5 A WEB SEMÂNTICA

A *Web* Semântica, segundo BERNERS-LEE (2001) - seu idealizador - “é uma visão – portanto, a ser implementada – de extensão da *Web* atual, que introduz uma estrutura e um significado para permitir a evolução de uma rede de documentos para uma rede de dados na qual toda a informação tem um significado bem definido para ser interpretada por computadores e humanos, aumentando assim a capacidade das máquinas de trabalhar em cooperação com as pessoas”.

FIGURA 7 - A VISÃO DE TIM BERNERS LEE DA WEB SEMÂNTICA



Fonte: http://www.ninebynine.net (2004)

Para Berners-Lee, a *Web* semântica está em construção e possibilitará o surgimento de uma “*web of trust*”, onde a confiabilidade das informações será integral.

Para atingir este nível se faz necessária a utilização de integração de tecnologias já existentes somadas a novas maneiras de trabalhar registros. Desta forma, partindo do *Unicode* e *Uris*, será feita a utilização da metalinguagem XML associada à sua gramática - o *XMLschema* – e aos NS que são os *Names Spaces* (locais na *Web* que concentram as definições sintáticas um recurso informacional).

Agrega-se a esses recursos a linguagem de metadados RDF e sua gramática RDF*schema*. Desta forma, as triplas podem identificar o recurso de maneira muito mais precisa.

Em seguida tem-se a agregação das ontologias que têm o papel de formalizar o conceito, sem que haja ambigüidade de interpretação.

Os agentes inteligentes, oriundos da Inteligência Artificial entram em cena como representantes da parte lógica e de inferência.

A esta altura da evolução atinge-se o patamar de se colocar à prova (*proof*) regras e mecanismos para os mais diversos fins.

As etapas *RDF+RDFschema, Ontology vocabulary, Logic e Proof,* serão viabilizadoras da utilização segura das Assinaturas Digitais (*Digital Signatures*).

Este patamar de evolução propiciará a denominada "*web of trust"* que tornará a recuperação da informação na *Web* segura e íntegra.

A *Web* Semântica surge como uma possível solução para a estruturação semântica dos dados na *Web*, viabilizando o processamento da informação por parte das máquinas. Está sendo desenvolvida por Tim Berners – Lee e sua equipe em um grupo de pesquisadores reunidos no *World Wide Web Consortium* ou W3C, que trabalha para melhorar, estender e padronizar o sistema.

5.1 O WORLD WIDE WEB CONSORTIUM (W3C)

A grande instituição de suporte a Web semântica é O W3C instalado nas dependências do *Massachussets Institut of Technology* (MIT) em Boston. Segundo documento do próprio W3C, seus objetivos e princípios operativos são:

a) acesso universal - o W3C define a *Web* como o universo de informação acessável por rede (disponível através de seu computador, telefone, televisão, ou geladeira conectada a uma rede...). Atualmente este universo beneficia a sociedade através da oferta de novas formas de comunicação entre humanos e oportunidades de compartilhamento de conhecimento. Um dos primeiros objetivos do W3C é tornar estes benefícios universais para todas as pessoas, independentemente de *hardware*, *software*, infraestrutura de rede, linguagem nativa, cultura, localização geográfica ou capacidades mentais ou físicas;

b) Web Semântica - as pessoas atualmente compartilham seu conhecimento na *Web* em uma linguagem destinada a outras pessoas. Na *Web* Semântica ("semântica" significa "relacionado a significado"), seremos capazes de nos expressar de modo a que os computadores possam interpretar a informação e fazer as trocas. Assim, será possível resolver problemas cotidianos tais como encontrar rapidamente informação médica, comentários sobre um filme, uma ordem de compra de um livro, etc;

c) confiança - a *Web* é um meio de colaboração e não apenas uma revista de leitura. Na realidade, o primeiro navegador para a *Web* era também um editor, apesar de a maioria das pessoas imaginar os navegadores com uma função principal de visualização e não interação. Para promover um ambiente mais colaborativo, torna-se necessário a existência de uma "Rede de Confiança" que garanta confidencialidade, passe confiança e torne possível às pessoas tomar responsabilidades por (ou ser responsáveis por) aquilo que está publicado na Rede. Estes objetivos orientam muito do trabalho no W3C's sobre assinaturas XML, mecanismos de anotação, autoridades de grupo, versões, etc;

d) interoperabilidade - há vinte anos as pessoas compravam *software* que funcionava apenas com algum outro *software* desde que fosse do mesmo vendedor. Atualmente as pessoas têm mais liberdade de escolha e corretamente esperam que os componentes de *software* sejam intercambiáveis. Também se espera que seja possível visualizar conteúdo existente na Rede com o software de sua preferência (navegador de PC gráfico, sintetizador de voz, navegador braille, telefone do carro, ...) A W3C, uma organização neutra a vendedores, promove interoperabilidade através do desenvolvimento e promoção de linguagens de computador abertas (não proprietárias) e protocolos que evitam uma fragmentação do mercado que existia no passado. Estes pontos são conseguidos através de consenso na indústria e encorajamento de uma discussão em fórum aberto;

e) evolução - o W3C visa a excelência técnica porém está ciente que o que conhecemos e necessitamos atualmente pode não ser insuficiente para a solução de problemas no futuro. Assim, luta para o desenvolvimento de uma Rede que possa facilmente evoluir para uma Rede ainda melhor, sem quebra de funcionalidades anteriores. Os princípios da simplicidade, modularidade, compatibilidade e extensibilidade orientam todo o seu desenvolvimento;

f) descentralização - a descentralização é um princípio de sistemas distribuídos modernos, incluindo-se aqui as próprias sociedades. Em um sistema centralizado toda a mensagem ou ação tem que passar por uma autoridade central, originando gargalos sempre que o tráfego

aumenta. Em conceito, limita-se então a quantidade de pontos centrais na Rede para reduzir a vulnerabilidade na Rede, como um todo. Flexibilidade é o elemento necessário para sistemas distribuídos e a vida e pulsar de toda a Internet, não apenas da Rede;

g) melhor multimídia - quem não gostaria de mais interatividade e uma melhor mídia na Rede, incluindo-se aqui imagens que podem alterar seu tamanho, som de qualidade, vídeo, efeitos tridimensionais e animação ? O processo de consenso no W3C não limita a criatividade de fornecimento de conteúdo ou significa visualizações de conteúdo aborrecidas. Através de seus membros o W3C escuta os usuários finais e trabalho com o objetivo de fornecer bases sólidas para o desenvolvimento de uma Melhor Multimídia através de linguagens como a linguagem *Scalable Vector Graphics* (SVG) e a *Synchronized Multimedia Integration Language* (SMIL).

5.2 ARQUITETURA DA WEB SEMÂNTICA

Na proposta de desenvolvimento da *Web* Semântica do W3C, é sugerida uma arquitetura de três camadas:

a) esquema: que estrutura os dados e define seu significado;
b) ontologia: que define as relações entre os dados;
c) lógica: que define mecanismos para fazer inferências sobre os dados.

5.2.1 Camada esquema

A camada esquema atua como uma gramática, garantindo que os dados nos documentos estejam bem estruturados e não apresentem ambigüidade de significado.

Para que se atinja o objetivo de estruturação dos dados disponíveis, se faz necessário, em primeiro lugar que estes dados sejam representados seguindo uma técnica adequada às características dos mesmos.

A partir da representação dos dados torna-se possível a representação do conhecimento. Essa representação do conhecimento é o primeiro passo no sentido de consolidar a *Web* Semântica.

Segundo FARIA e GIRARDI (2004, p.2) se faz necessário a satisfação de três condições para que haja a representação do conhecimento:

a) interoperabilidade estrutural, que provê a representação para modelos de dados distintos, permitindo especificar tipos e possíveis valores para cada forma de representação;
b) interoperabilidade sintática, responsável por regras precisas para promover o intercâmbio dos dados na *Web*;
c) interoperabilidade semântica, que possibilita a compreensão dos dados e suas associações com outros dados.

Afim de que a camada esquema possa atuar como uma gramática de definição faz-se uso de alguns artifícios tecnológicos tais como a linguagem XML, a meta-linguagem RDF e identificadores universais URI, etc.

A linguagem XML utiliza *tags* para definir o início e o fim do texto marcado como unidade ou elemento de informação, permitindo tratar cada unidade de informação como um objeto ou uma entidade ao qual se pode atribuir características específicas.

A estrutura do XML é formada basicamente por três elementos que se completam para determinar o armazenamento do conteúdo e sua apresentação:

a) DTD (*Document Type Definition*) ou Esquema XML;
b) CSS (*Cascading Style Sheets*) / XSL (*eXtensible Style Language*);
c) Conteúdo XML.

O elemento DTD faz a verificação da correção do documento na sua forma sintática e define os elementos que constituem a estrutura do documento.

O elemento CSS é o encarregado de especificar de que forma o documento irá ser apresentado.

O conteúdo XML é um documento XML.

5.2.1.1 HTML versus XML na representação do conhecimento

Conforme abordado anteriormente, a linguagem HTML é responsável pela marcação de texto, enquanto que a XML vai além, marcando conteúdos.

Esta diferença tem implicações poderosas quanto se trata da recuperação da informação com qualidade.

A Figura 8 ilustra a maneira como as linguagens codificam a informação.

Enquanto que XML distribui a informação entre *tags* bem específicas – como por exemplo <modelo> Pentium IV</modelo>, a HTML dispõe toda a informação entre uma única *tag* <body>.

É possível visualizar que a opção XML é mais indicada para a recuperação da informação, pois possibilita mais precisão na escolha de um descritor que levará a uma maior precisão dos resultados retornados.

FIGURA 8 – EXEMPLO DE CODIFICAÇÃO EM XML E HTML

```
<?xml version="1.0"?>
<!-- Microcomputador -->
<microcomputador>
<modelo> Pentium IV </modelo>
<velocidade>1.5 GHz </velocidade>
<ram> 256MB </ram>
<monitor> 17 polegadas </monitor>
<teclado> Sim </teclado>
<mouse> Sim </mouse>
<estabilizador>Sim</estabilizador>
<impressora>Não </impressora>
</microcomputador>
```

Código XML

```
<html>
<body>
Microcomputador
Pentium IV,
1.5GHz,
256MB de RAM,
Monitor 17 polegadas,
Mouse,
Teclado,
Estabilizador.
</body>
</html>
```

Código HTML

FONTE: FARIA e GIRARDI (2004)

5.2.1.2 Atendimento das condições para a representação do conhecimento

Apesar de a linguagem XML ser mais completa para a apresentação de conteúdos, ainda não atende apenas a duas das condições para a representação do conhecimento citadas em seção anterior. Ou seja, a interoperalidade sintática e a interoperalidade estrutural, ambas atendidas através do DTD ou Esquema XML.

A interoperalidade semântica não é atendida pois a XML estrutura os dados sem porém dar-lhes significado.

5.2.1.3 *Uniform Resource Idenrifier (*URI) e RDF

Um URI (*Uniform Resource Identifier*) é um identificador de recursos disponíveis na *Web*. Através de uma série de caracteres iniciados por “http:” ou “ftp:”, é possível localizar principalmente recursos informacionais na *Web*.

Dentre os diversos tipos de URI, o mais conhecido é o URL (*Uniform Resource Locator*) utilizado para identificar o local específico de um determinado recurso.

Um vez que o URI podem ser utilizados para identificar qualquer recurso, torna-se possível fazer qualquer distinção necessária entre termos, por exemplo entre homônimos. Por outro lado o URI também pode ser utilizado para agrupar termos diferentes que tenham o mesmo significado.

Desta forma, para cada significado específico existirá uma tripla que o identifica. Este é um requisito denominado unicidade, fundamental para a constituição de uma rede de informação a partir destes significados e para o processamento pelas máquinas.

O RDF providencia um mecanismo simples de representação de conhecimento dos recursos da Web e consegue satisfazer as três condições para a interoperabilidade sintática, a interoperabilidade estrutural e a interoperabilidade semântica. A primeira através do Esquema RDF (regras precisas), a segunda através da representação por meio de triplas e a terceira – a semântica – através do significado único atribuído aos dados de cada tripla.

5.2.2 Camada Ontologia

Apesar do uso das triplas garantirem uma definição única para os conceitos, pode existir a condição em que um mesmo conceito pode ser expresso de forma e linguagem diferentes. Isto pode ocorrer devido a padronizações criadas para sistemas particulares.

A função da camada ontologia é identificar uma relação entre os conceitos apresentados em documentos distintos para viabilizar o processamento.

5.2.3 Camada Lógica

Na seção anterior demonstrou-se que a camada ontologia deve identificar a relação entre conceitos. Esta identificação inequívoca permite a atuação da próxima camada denomina de lógica.

Na camada lógica ocorre a ação de regras de inferências utilizadas pelos agentes para processar a informação. Essas regras permitem que os agentes atuem de maneira inteligente sobre os conteúdos e seus significados.

5.3 AGENTES INTELIGENTES E RECUPERAÇÃO DA INFORMAÇÃO

"Um agente é um sistema computacional que está situado em um ambiente e que é capaz de atuar de forma autônoma neste ambiente com intenção de atingir os objetivos de seu desenvolvedor" (WOODRIDGE, 1999).

Algumas características dos agentes são: autonomia, reatividade, flexibilidade capacidade de aprendizagem. Estas características entre outras permitem aos agentes "compreender" e fazer inferências sobre as relações existentes entre os conteúdos.

A atuação dos agentes eleva a autenticidade e confiabilidade das fontes de informação a um patamar ainda não esclarecido completamente, porém é certo que de qualidade extremamente superior se comparada a encontrada atualmente.

É esperado que os agentes possam cooperarem entre si, compartilhando ontologias para expandir suas capacidades.

5.4 EVOLUÇÃO DA WEB SEMÂNTICA

Da mesma forma que aconteceu com a *Web*, a adoção da *Web* semântica se dará de forma gradual. Isto deve-se em parte à disponibilização da tecnologia e principalmente à criação de ontologias e à interconexão entre elas.

Sob uma perspective técnica (SEMAVIEW, 2002), a *Web* semântica evoluirá em uma série de estágios considerados a seguir:

a) estágio 1 - ilhas semânticas

Este é estágio atual da *Web* semântica. Nele, ilhas de conteúdo e aplicativos semânticos estão sendo desenvolvidos em ambientes corporativos. Estas ilhas podem ser utilizadas internamente nas empresas ou entre parceiros de negócios num cenário *business to business* (B2B).

b) estágio 2 - ilhas conectadas

Neste estágio, a quantidade de ilhas crescerá exponencialmente possibilitando interseções entre elas na medida em que o conteúdo é disponibilizado através da internet.

Softwares autônomos – agentes – atuarão na Internet fazendo uso da vasta coleção de conteúdo semanticamente disponibilizado. Estes agentes buscaram informações com elevadíssimo grau de precisão. Eles terão capacidade de interagir com a rede sem a anuência e participação de humanos.

c) estágio 3 - Inteligência Artificial

Em um estágio mais distante, será possível a utilização maciça da inteligência artificial para criar e usar a grande quantidade de metadados disponíveis.

Este estágio está ainda longe de ser alcançado e ninguém sabe ao certo o que ocorrerá. Porém, é certo que os agentes evoluirão significativamente no sentido de capacidade de inferências e inteligência.

5.5 WEB SEMÂNTICA E RECUPERAÇÃO ONLINE DA INFORMAÇÃO

A partir do quadro teórico desenvolvido nas sessões anteriores, torna-se possível vislumbrar o impacto da *Web* semântica em vários setores, sejam eles ligados à produção industrial, serviços, educação, etc.

Processos de recuperação *online* da informação, sejam eles considerados atividade-fim ou atividade-meio serão os maiores beneficiados com os avanços propostos pela *Web* semântica.

Mecanismos de busca capazes de utilizarem documentos marcados semanticamente, isto é, através do uso de XML, RDF e Ontologias retornarão resultados muito mais precisos.

Além dessas tecnologias, deve-se levar em consideração a importante contribuição dos agentes inteligentes, capazes de tomarem decisões e atuarem de maneira autônoma na busca de informações previamente indicadas pelo usuário.

Conforme discussão anterior, a *Web* semântica se encontra no seu primeiro estágio, o de ilhas semânticas. Assim nota-se apenas produções isoladas de "aplicativos semânticos". Este têm atuação limitada a uma área do conhecimento ou a uma instituição.

Dentre os aplicativos, destaca-se a iniciativa de produzir mecanismos de busca que possam interagir com conteúdos semanticamente marcados.

Uma destas tentativas chama-se *Swoogle*[1], um mecanismo de busca em construção pela Universidade de Maryland nos EUA.

Swoogle é um projeto de pesquisa do *Computer Science and Electrical Engineering Department of the University of Maryland, Baltimore County.*

O *Swoogle* é um sistema de recuperação de informação dedicado à *Web* semântica, ou seja, à recuperação de documentos codificados através de RDF, OWL e XML. Este sistema extrai metadados de cada documento e verifica as relações entre eles.

Além disso o *Swoogle* faz um *rank* de ontologias que mede a importância do *Semantic Web Document* (SWD).

Atualmente o *Swoogle* tem uma base de dados contendo 308.297 (trezentos e oito mil, duzentos e noventa e sete) documentos semânticos (SWDs) que contém 43.104.786 (quarenta e três milhões, cento e quatro mil, setecentos e oitenta e três) Triplas do tipo A(O,V). (Dados referentes a 28 de novembro de 2004).

Usando o *Swoogle*, pode-se encontrar todos os documentos que usam determinado atributo, objeto ou valor.

5.5.1 *Interface* do *Swoogle*

Visualmente, a *interface* é semelhante ao *Google*, porém as similaridades se encerram neste ponto.

[1] http://swoogle.umbc.edu

FIGURA 9 – INTERFACE SWOOGLE

FONTE: http://swoogle.umbc.edu

Apesar da interface, os propósitos do *Swoogle* são mais complexos. Ele é destinado à usuários mais especializados, requer um conhecimento básico de ontologias, atributos, etc.

A definição de *Web* semântica considera a operação direta entre máquinas. O *Swoogle* é um repositório de ontologias que poderão ser usadas por agentes inteligentes em suas ações independentes da ação humana.

O *Swoogle*, bem como os próximos mecanismos que certamente surgirão, requer um usuário diferenciado dos que atualmente fazem uso dos mecanismos comuns no mercado. Os resultados apresentados pelo *Swoogle* podem parecer totalmente estranhos para um usuário não iniciado em ontologias e RDF por exemplo.

O usuário padrão dos mecanismos semânticos de busca utilizará a ferramenta como complemento a um desenvolvimento próprio de ontologias, *NamesSpaces*, etc. Certamente será um usuário diferenciado, muito provavelmente especializado em uma área do conhecimento.

5.5.2 Outras Aplicações para *Web* Semântica

A *Web* semântica terá impacto não apenas na recuperação da informação. Talvez esta nem seja a área de maior intervenção.

Existem projetos sendo desenvolvidos em diversas outras áreas tais como, recursos humanos, hotelaria, economia e finanças, *e-learning*, aplicativos para uso pessoal, etc.

Neste tipo de aplicações, os agente inteligentes terão papel fundamental pois espera-se que eles possam assumir tarefas humanas, tais como marcar uma consulta como o médico sem que o paciente precise informar manualmente detalhes tais como a sua disponibilidade na agenda bem como a agenda do médico.

Os agentes poderão, através dos recursos da *Web* semântica, confrontar as duas agendas e escolher o horário vago em comum aos dois interessados.

5.5.3 Implicações Negativas no uso da *Web* Semântica

As facilidades oferecidas pela *Web* semântica – como qualquer outra tecnologia – apresentam suas implicações negativas. E como sempre, estas implicações não são inerentes à tecnologia em si, mas ao uso que se faz dela.

Por exemplo, no aplicativo acima sugerido, capaz de agendar uma consulta, fica clara a opção pela redução da privacidade de ambas as partes. Para que o agente possa escolher a data livre e precisará “invadir” a privacidade do médico. A ação será recíproca com relação à privacidade do paciente.

6. CONSIDERAÇÕES FINAIS

A *Web* semântica está em processo de formação, ainda restrita normalmente a ambientes acadêmicos com algumas iniciativas privadas.

Toda nova tecnologia que surge, tende a ser adotada pelas instituições quase que imediatamente, considerando principalmente o fator diferencial competitivo. Nesta corrida, algumas empresas tentarão apresentar "soluções semânticas" para seus clientes.

Porém, para que a *Web* semântica possa ser implementada em soluções verdadeiramente semânticas - sejam quais forem – existe uma condição primordial que é a criação de um número suficientemente grande de ontologias na área em questão. Só a partir do momento em que determinado conceito está devidamente formalizado em uma série de ontologias é que se torna possível a localização e troca dos recursos sem que haja perigo de ambigüidades e incongruências, no caso da recuperação da informação, imprecisão.

Já se pode inferir como segurança que o impacto na *Web* atual será bastante significativo. Não haverá uma suplantação, até mesmo porque o próprio Berners-Lee argumenta que a *Web* semântica é uma extensão da *Web* atual.

Na área da Ciência da Informação é possível vislumbrar uma nova era de qualidade e precisão das informações recuperadas pelos mecanismos com base semântica.

Dentro deste contexto, o profissional da informação, em especial o Gestor da Informação, terá o papel fundamental de acompanhar as iniciativas relativas à sua atuação. Desta maneira estará apto a fornecer informação útil e precisa para que esta cumpra sua finalidade principal: dar suporte à tomada de decisão de indivíduos e instituições.

REFERÊNCIAS

AMARAL, Marcelo. **XML em 10 pontos**. Disponível em: < http://paginas.terra.com.br/informatica/mja/W3C/XML-in-10-points.pt-BR.html > Acesso em: 28 nov. 2004.

ABBAGNANO, N. **Dicionário de filosofia**. São Paulo: Ed. Mestre Jou, 1970.

BERNERS – LEE, Tim. ***Semantic Web Roadmap.*** Disponível em: <http://www.w3c.org/DesignIssues/Semantic.html> Acesso em: 02 set. 2004.

BERNERS – LEE, Tim et al. ***Semantic Web Development Proposal.*** Disponível em < http://www.w3c.org/2001/sw> . Acesso em 02 set. 2004

BERTALANFFY, L. **International Society for the Systems Sciences**. Disponível em: < http://www.isss.org/lumLVB.htm>. Acesso em: 14 mai. 2004.

BIOLCHINI, Jorge. **Semântica e cognição em bases de conhecimento: do vocabulário controlado à ontologia.** DataGramaZero – Revista de Ciência da Informação – v.2 n.5 out. 2001.

FARIA, Carla; GIRARDI, Rosario. **Uma análise da web semântica e suas implicações no acesso à informação**. Disponível em: <http://maae.deinf.ufma.br/Ensino/IA>. Acesso em: 25 out. 2004.

FERREIRA, Aurélio Buarque de Holanda. **Novo Aurélio Século XXI**: o dicionário da língua portuguesa. 3. ed. rev. ampl. Rio de Janeiro: Nova Fronteira, 1999.

HENDLER, James. ***Agents and the Semantic Web***. Disponível em: <www.computer.org/internet > Acesso em: 21 jul. 2004.

JASPER, R; USCHOLD, M. ***Enabling task-centered knowledge support through semantic metadata***. In: Semantic Web Technology. MIT Press. 2001.

KASABOV, Nikola. ***Foundations of neural networks, fuzzy systems, and knowledge engineering***. Londres: MIT Press. 1996.

LAGOZE, K**. keeping Dublin Core simple**. Disponível em: < http://www.dlib.org/dlib/january01/lagoze/01lagoze.html>. Acesso em: 28 nov. 2004.

LASSILA, O; SWICK, R. ***Resource Description Framework (RDF) Model and Syntax Specification.*** Disponível em: <ttp://www.w3c.org/TR/1999/REC-rdf-syntax-19990222> . Acesso em 06 ago. 2004.

LANCASTER; WARNER. ***Estratégia de busca na recuperação da informação***. Disponível em: www.ibict.br/cienciadainformacao/include/getdoc.php?id= 477&article= 191&mode=pdf. Acesso em: 15 abr. 2004.

LANDOW, G. ***La convergencia de la teoría crítica contemporánea y la tecnología***. Disponível em: http://www.ucm.es/info/especulo/numero2/landowhi.htm>. Acesso em: 12 mai. 2004.

MILSTEAD, J.; FELDMAN, S. **Metadata: cataloging by any other name**. Disponível em: <www.onlineinc.com/onlinemag/metadata>. Acesso em: 15 jul. 2004.

MORIN, E. **O Método: o conhecimento do conhecimento**. Lisboa: Publicações Europa-América Ltda, 1986.

OLIVEIRA, Rosa. **Web semântica: o novo desafio para os profissionais da informação**. Disponível em: < http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/124.a.pdf> Acesso em: 01 jun. 2003**.**

SCHNEIDERMAN, B. ***Hypertext***. Disponível em: <http://www.cs.umd.edu/~ben/>. Acesso em: 26 set. 2004.

SEMAVIEW. ***The semantic web illustrated***. Disponível em: <http://www.semaview.com> Acesso em: ago. 2004.

SETZER, V.W. **Dado, informação, conhecimento e competência**. Disponível em :<http://www.ime.usp.br/~vwsetzer/dado-info.html> Acesso em: 04 set 2004.

SMITH, R. ***What's required in knowledge technologies***. Disponível em: http://www.gca.org/attend/2001_conferences/kt_2001/default.htm. Acesso em: 25 jun. 2004.

SOUZA, R.; ALVARENGA, l. A web semântica e suas contribuições para a ciência da informação. Ci.inf, Brasília, v.33, n. 1, p. 132-141, jan./abr.2004.

TAKAHASHI, T. (Org.). **Sociedade da informação no Brasil**: livro verde. Brasília: Ministério da Ciência e Tecnologia, 2000.

TRIPPE, Bill. ***Taxionomies and topic maps*****.** Econtent Magazine. Disponível em <http://www.ecmag.net/Magazine/Features/trippe8_01.html> Acesso em: 04 mar. 2004.

WEIBEL, Stuart. ***Metadata: the foundations of resource description***. D-Lib Magazine. Jul. 1995.

WOODRIDGE, M. ***Intelligent agents***. New York: MIT Press, 1999.

World Wide Web Consortium (W3C). Disponível em <http://www.w3c.org> Acesso em: 04 mar. 2004.

ANEXO – *THE SEMANTIC WEB ILUSTRATED*