

Web Semântica: ontologias como ferramentas de representação do conhecimento

Maria Elisa Valentim Pickler

Universidade Estadual de Londrina;

Bolsista de IC/PROIC

A Web Semântica é uma extensão da Web que acrescenta semântica ao atual formato de representação de dados. Para isso foram propostas diversas tecnologias, dentre essas a criação de ontologias, visando atribuir sentido e significado ao conteúdo dos documentos, atuando como ferramenta de representação do conhecimento. Na presente pesquisa, procurou-se verificar se as ontologias seriam utilizadas apenas para representar o assunto de uma página Web através de termos contextualizados ou se tentariam controlar o vocabulário da Web como um Tesouro.

Palavras-chave: *Web Semântica; Ontologias; Tesouros.*

Semantic Web: ontologies as tools of knowledge representation

The Semantic Web is an extension of the Web that adds semantics to the current format of data representation. For this, several technologies were proposed, among them the creation of ontologies, aiming at attributing sense and meaning to the content of documents, acting as tool for knowledge representation. In the present research, it was attempted to verify whether the ontologies would be used only to represent the subject of a Web page through terms in context or rather, they would try to control the vocabulary of the Web as a thesaurus.

Key-words: *Semantic Web; Ontologies; Thesaurus.*

Recebido em 21.08.2006 Aceito em 14.03.2007

Introdução

O desenvolvimento da *Web* possibilitou o surgimento de um novo meio de interação e comunicação em sociedade que, além de absorver todas as mídias anteriores, permitiu o crescimento explosivo da quantidade de informação e conhecimento disponíveis e acessíveis em rede a todo o mundo. Como conseqüência, houve a necessidade de ferramentas capazes de encontrar entre inúmeros dados irrelevantes, uma informação precisa. Para tanto, foram criados os mecanismos de busca.

A clássica organização do conhecimento realizada por áreas como a Biblioteconomia e a Documentação é fundamentada na classificação de conceitos proposta por Aristóteles, formando diversas subcategorias para que todo e qualquer informação ou conhecimento produzido estivesse contemplado em algumas dessas classes/categorias pré-estabelecidas.

A organização do conhecimento realizada através da adoção de terminologias padronizadas e o controle da linguagem possibilitada pela utilização dos *tesauros*, eliminaram a polissemia, não permitindo ambigüidades nas linguagens documentárias. No entanto, no ciberespaço¹, esse controle de terminologia da linguagem dos documentos já não é possível (ou necessário) considerando que sua multiplicidade permite a polissemia, característica intrínseca à própria linguagem natural, base da indexação na rede. Monteiro (2006, p. 35) ressalta:

No ciberespaço não há centro de significância estruturado, hierarquizado, linear, ou instrumentos de organização do conhecimento que reproduzem o modelo de significância, sentido único e referência fixa. A multiplicidade e as várias possibilidades de tratamento da informação no ciberespaço ilustram que não há uma maneira correta de organização do conhecimento que deva ser empregada e reproduzida pelos outros sistemas para tentar atingir a universalidade [...].

Segundo a autora, a característica do ciberespaço é justamente essa ausência de classificação do conhecimento; o ciberespaço permite a virtualização da linguagem que a organização formal do conhecimento impede ao retirar a sua ambigüidade.

Assim sendo, para recuperar informações disponíveis na *Web*, os mecanismos de busca valem-se da indexação de palavras (em linguagem natural) presentes nos próprios documentos, embora cada ferramenta opere segundo princípios por vezes divergentes.

A polissemia da linguagem natural apresenta-se como um obstáculo na recuperação de informações realizada pelos mecanismos de busca da *Web*. Uma vez que a *Web* contempla um número inestimável de documentos de todas as áreas possíveis do conhecimento humano, ao realizar uma busca por determinado termo, os mecanismos de busca

1 Ciberespaço, para Monteiro e Pickler (2006), é o ambiente virtual possibilitado pelas tecnologias da internet. Um novo espaço de comunicação, representação e interação em sociedade. Já a *Web* pode ser entendida como um ambiente multimídia da Internet, que disponibiliza o conteúdo em formato de hipertexto. Nesse sentido, a *Web* é o principal lugar do ciberespaço, seu principal edifício, podendo tomá-la como o centro de todas as possibilidades de interfaces.

trarão, como resultados, todos os *sites* (que constam em sua base de dados) que apresentem em seu conteúdo as palavras buscadas. Cabe ao usuário (humano) consultar cada *site* obtido a fim de verificar se o conteúdo do documento recuperado é pertinente ao seu contexto e sua necessidade, ou seja, se a palavra buscada aparece nos resultados com o mesmo sentido e significado.

Nessa direção, Souza e Alvarenga (2004) consideram que a dificuldade em determinar os contextos informacionais tem como conseqüência a impossibilidade de se identificar de forma precisa a atinência dos documentos. Além disso, a ênfase das tecnologias e linguagens utilizadas nas páginas da *Web* tradicional focaliza os aspectos de exibição e apresentação dos dados, de uma forma que a informação seja descrita pobremente e pouco passível de ser consumida concomitantemente por máquinas e seres humanos. A partir disso é que surge a proposta da *Web Semântica*.

A *Web Semântica* foi proposta por Berners-Lee para tentar melhorar (ou mesmo otimizar) as pesquisas realizadas na *Web*. É uma nova ferramenta de busca de informações no ciberespaço, uma extensão da *Web*, acrescentando semântica ao atual formato de representação de dados. Uma ferramenta inteligente, que trabalha através de associação e dedução. Sob essa perspectiva, a tarefa de verificar o assunto do documento ficaria a cargo das máquinas, poupando tempo e trabalho a quem realizasse uma busca.

Nesse contexto, foi proposta uma série de tecnologias necessárias para que as máquinas sejam dotadas de ferramentas inteligentes e, assim, sendo capaz de raciocinar, inferindo (deduzindo) o conteúdo dos documentos armazenados em seu banco de dados.

Uma dessas tecnologias necessárias é a elaboração e utilização de *ontologias*, não no sentido da Filosofia, como estudo do ser e da existência, mas como uma forma de atribuir *sentido e significado* a determinados termos, em dados contextos, em busca de atribuir semântica ao conteúdo dos documentos, atuando como ferramenta de representação do conhecimento.

Sendo assim, o que consideramos um problema a ser estudado é de que maneira as ontologias proporcionarão a organização do (quase) infinito conteúdo na caótica *Web*. Se as ontologias seriam utilizadas apenas para representar o assunto de uma página *Web* através de termos contextualizados ou se tentariam controlar o vocabulário da *Web* como um tesouro que estabelece uma linguagem controlada.

A ontologia seria uma ferramenta/linguagem que permite a instauração de sentido, mas o sentido atualizado de acordo com o contexto do leitor, minimizando a polissemia para quem busca determinado assunto, mas sem excluí-la do ciberespaço, ou diz respeito aos conceitos do tesouro, quer seja, uma rubrica autorizada à organização e recuperação do conhecimento, operando apenas na instauração de um significado.

Nesse sentido, o objetivo geral da presente pesquisa consistiu em analisar a literatura científica da área e/ou áreas afins para definir e

caracterizar a *Web Semântica*. Para isso, foi necessário estabelecer alguns objetivos específicos, tais como: definir e caracterizar a *Web* atual; apresentar alguns tipos de mecanismos de busca disponíveis na *Web*, como os diretórios, motores de busca e metabuscadores; identificar, na literatura, as ontologias; estudar, na literatura, a *Web Semântica*.

A presente pesquisa procurou verificar como se dá a recuperação de informação e conhecimento na *Web* através dos mecanismos de busca, para basear o estudo da extensão da *Web* que foi proposta e ainda não foi desenvolvida: a *Web Semântica*.

A *Web Semântica* objetiva aprimorar o serviço de busca e recuperação de dados na *Web*, embora ainda não exista. Por esse motivo, esta pesquisa adquire o caráter de Pesquisa Documental, baseada apenas na literatura, uma vez que objeto da pesquisa (*a Web Semântica*) está, por ora, em construção. Como consequência, a pesquisa não teve um *corpus* já que, como dito anteriormente, o objeto específico ainda não foi desenvolvido.

Assim sendo, esta pesquisa iniciou-se com o levantamento bibliográfico da literatura da área e áreas afins que contemplam o objeto. Toda a pesquisa foi desenvolvida em torno da revisão de literatura científica obtida no levantamento bibliográfico, tratando-se, portanto, de uma pesquisa teórica.

A Web semântica

A *World Wide Web* foi criada por Tim Berners-Lee entre 1989 e 1991, com base nas idéias originadas de trabalhos anteriores sobre hipertexto realizados por Bush na década de 40 e por Ted Nelson na década de 60.

A *Web* surgiu com a visão de que seria um espaço onde a informação poderia adquirir um significado bem definido, de forma que facilitasse a cooperação e a comunicação entre as pessoas e os agentes computacionais, conforme observa Cunha (2002). Entretanto, a *Web* voltou-se mais para a comunicação entre os humanos.

A *Web* atual é denominada por Breitman (2005) de *Web Sintática*, na qual os computadores fazem apenas a apresentação da informação, enquanto o processo de interpretação fica a cargo dos seres humanos, já que isso exige um grande esforço para avaliar, classificar e selecionar informações e conhecimentos de interesse.

Embora a *Web* tenha sido projetada com a finalidade de possibilitar o fácil acesso, o intercâmbio e a recuperação de informações, Souza e Alvarenga (2004) afirmam que ela foi implementada de forma descentralizada e quase anárquica, cresceu de maneira exponencial e se apresenta atualmente como um imenso repositório de documentos que deixa muito a desejar quando se trata de recuperação de conteúdo relevante. Os autores fundamentam suas críticas afirmando que:

Não há nenhuma estratégia abrangente e satisfatória para a indexação dos documentos nela contidos, e a recuperação das

informações, possível por meio dos motores de busca (search engines), é baseada primariamente em palavras-chave contidas no texto dos documentos originais, o que é muito pouco eficaz. (SOUZA e ALVARENGA, 2004, p. 133).

Nesse sentido, cabe notar que os mecanismos de busca são extremamente ricos em quantidade de *sites* indexados em suas bases de dados, embora sequer os melhores deles consigam abranger a totalidade de conteúdo disponível na *Web*.

Apesar de a quantidade de informações recuperadas pelos mecanismos de busca ser massiva, apenas uma parte da *Web* é pesquisada, enquanto uma parte considerável do conteúdo fica inacessível através dos buscadores (*Web Oculta*).

Além disso, a dificuldade de se encontrar informações relevantes através dos buscadores é mascarada por suas interfaces aparentemente amigáveis. Assim, apesar da grande quantidade de informação na *Web* e dos mecanismos disponíveis para buscá-las, o usuário muitas vezes fica frustrado com os resultados insatisfatórios que lhe são retornados.

Nesse sentido, a *Web Semântica* visa, justamente, melhorar a satisfação do usuário no momento da busca, retornando-lhe as informações adequadas às suas necessidades.

Contraopondo essa *Web Sintática*, surge a *Web Semântica*, através da qual se buscam mecanismos que capturem o significado das páginas, criando um ambiente no qual os computadores possam processar e relacionar conteúdos provenientes de várias fontes. Para que isso se torne possível, é necessário embutir semântica na estrutura dos documentos disponíveis na *Web* (BREITMAN, 2005).

A palavra *semântica*, segundo Guirald (1980), é formada do grego *sêmainô* (significar), derivado de *sema* (sinal), que corresponde a sentido. Para o autor, é semântica tudo o que se refere ao sentido de um sinal de comunicação e tudo o que se refere às palavras.

Dessa forma, Chateaubriand (1998) afirma que a semântica é a disciplina que estuda a relação da linguagem com a realidade, ou seja, a relação linguagem-mundo. Exemplificando, o autor ressalta que o fato de que a palavra casa se refira, na realidade, a casas e não a elefantes é um fato semântico.

Se as palavras codificam um sentido de várias maneiras, podemos entender que a semântica é o estudo da função das palavras, função essa de transmitir um sentido e um significado relativos a um conteúdo. Sendo assim, percebemos que, se a intenção inicial da *Web Semântica* é justamente acrescentar semântica ao conteúdo da *Web*, essa semântica servirá para determinar o sentido de um termo no contexto de determinado documento.

A *Web Semântica* não é, propriamente, uma outra *Web*. Oliveira (2002) afirma que Tim Berners Lee a caracteriza como uma extensão da *Web* atual que apresentará uma estrutura que possibilite a compreensão e o gerenciamento dos conteúdos armazenados na *Web* independentemente da forma em que esses se apresentam (texto, som, imagem) a partir da

avaliação semântica desses conteúdos, e através de agentes coletores de conteúdos advindos de fontes diversas capazes de processar as informações e permutar resultados com outros programas.

A *Web Semântica*, assim como a atual *Web (Web Sintática)*, será tão descentralizada quanto possível e deverá manter a responsabilidade exigida pela descentralização, procurando alcançar o ideal de consistência de interconexões, permitindo, porém, seu crescimento exponencial. Com efeito, a extensão da *Web*, na esteira de Dziekaniak e Kirinus (2004), surge como uma possível solução para a estruturação dos dados nesse meio, viabilizando o processamento da informação por parte das máquinas e melhorando a recuperação de informações.

Enquanto a *Web Sintática* foi desenvolvida para ser entendida apenas pelos usuários, a *Web Semântica* está sendo projetada para ser compreendida pelas máquinas, na forma de *agentes* computacionais que serão capazes de operar eficientemente sobre as informações, podendo até entender (inferir) seus significados. Assim, esses agentes auxiliarão os usuários em suas diversas operações na *Web*.

A proposta da *Web Semântica* é estruturar os dados contidos nos *sites* de uma forma que o próprio sistema de busca identifique seu assunto e conteúdo e para isso seria preciso embutir *semântica* na estrutura dos dados.

Assim, ao buscar um termo polissêmico por um de seus significados, o usuário teria como resultados apenas os *sites* nos quais aquele termo aparece relacionado ao sentido buscado por ele.

Nesse contexto, podemos perceber que a *Web Semântica* visa facilitar e melhorar a recuperação de informação relevante, já que a própria máquina, dotada de ferramentas inteligentes, funcionaria por associação e dedução automática para identificar (inferir) o conteúdo de um *site* antes de trazê-lo ao usuário como resultado de uma pesquisa.

Souza e Alvarenga (2004) observam que, para atingir os propósitos da *Web Semântica*, é necessária uma padronização de tecnologias, de linguagens e de metadados descritivos, de forma que todos os usuários da *Web* obedeçam a determinadas regras comuns e compartilhadas sobre como armazenar dados e descrever a informação armazenada de forma que esta possa ser consumida por outros usuários (humanos ou não), de uma maneira automática e não ambígua.

Sendo assim, a partir da existência da infra-estrutura tecnológica comum da Internet, o primeiro passo para esse objetivo é a criação de padrões para descrição de dados e de uma linguagem que permita a construção e codificação de significados compartilhados.

Revisitando Souza e Alvarenga (2004), o projeto da *Web Semântica*, em sua essência, é criar e implantar padrões (*standards*) tecnológicos para permitir esse panorama, que não somente facilite a troca de dados entre os agentes pessoais, mas estabeleça uma língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

Dziekaniak e Kirinus (2004) acreditam que, para a *Web Semântica* se tornar possível, os computadores necessitam ter acesso a coleções

estruturadas de informações (dados e metadados) e de conjuntos de regras de inferência que ajudem no processo de dedução automática para que seja administrado o raciocínio automatizado, ou seja, a representação do conhecimento. Essas regras, para as autoras, são especificadas por meio de *ontologias* que permitem representar explicitamente a semântica dos dados.

Através das ontologias será possível elaborar uma enorme rede de conhecimento humano, complementando o processamento da máquina e melhorando qualitativamente o nível de serviços na *Web*, sobretudo os serviços de busca e recuperação de dados.

Nesse sentido, Hendler *apud* Dziekaniak e Kirinus (2004) afirma que a *Web Semântica* pode ser considerada como a composição de um grande número de pequenos componentes ontológicos que se apontam entre si. Dessa forma, companhias, universidades, agências governamentais e grupos de interesses específicos procurarão ter seus recursos *Web* ligados a um conteúdo ontológico, uma vez que ferramentas poderosas serão disponibilizadas para intercambiar e processar essas informações entre aplicações *Web*.

Ontologias

A palavra ontologia é derivada do grego: *onto* (ser) e *logos* (palavra). O termo ontologia foi cunhado em 1613 por Rudolf Goclenius e, aparentemente, de forma independente por Jacob Lorhard. (MOREIRA *et al*, 2004).

De acordo com Souza e Alvarenga (2004), na filosofia, a ontologia é uma teoria sobre a natureza da existência. Chateaubriand (1998, p. 12) afirma que:

Em princípio, pode-se dizer que ontologia é uma teoria do real, uma teoria do ser. Essa é uma concepção tradicional que afirma, em geral, que a ontologia envolve alguma categorização muito ampla da realidade. Isto é, para se fazer uma teoria ontológica, uma das primeiras coisas a fazer, e é o que foi feito na tradição filosófica, é categorizar de alguma maneira a realidade em certas grandes distribuições do ser.

Para o autor, uma das tendências mais fortes do século XX é fazer a ontologia, de certa forma, derivar da semântica. A semântica, para o autor, tem a ver com a relação entre linguagem e realidade e é a partir dessa idéia que, evidentemente, ontologias e semântica se conectam. Com efeito, se a semântica tem a ver com a relação entre a linguagem e a realidade e a ontologia é o estudo da estrutura geral da realidade, então necessariamente haverá relações entre elas.

Breitman (2005) afirma que o vocábulo ontologia foi introduzido no estudo da Filosofia para distinguir o estudo do ser e o estudo dos vários tipos de seres vivos existentes no mundo natural, tendo o objetivo de fornecer sistemas de categorização para organizar a realidade. No

entanto, o uso do termo ontologia tornou-se freqüente na ciência da computação no início dos anos 90, em projetos para organização de grandes bases de conhecimento, como ressaltam Moreira *et al* (2004). Para os autores, um dos principais objetivos do uso de ontologias na Ciência da Computação é a construção de bases de conhecimento interoperáveis e melhor estruturadas.

Assim, no contexto da *Web e da Inteligência Artificial*, o termo ontologia foi adaptado e, para os profissionais dessas áreas, uma ontologia é um documento ou um arquivo que define formalmente as relações entre termos e conceitos, mantendo, nesse sentido, *semelhanças com os tesouros* utilizados para definição de vocabulários controlados (SOUZA e ALVARENGA, 2004).

Na concepção de Breitman (2005, p. 7):

Ontologias são especificações formais e explícitas de conceitualizações compartilhadas. Ontologias são modelos conceituais que capturam e explicitam o vocabulário utilizado nas aplicações semânticas. Servem como base para garantir uma comunicação livre de ambigüidades. Ontologias serão a língua franca da Web Semântica.

Dessa forma, as ontologias fornecerão o vocabulário necessário para a comunicação entre os agentes e as páginas da *Web*, definindo as relações entre os conceitos, como salientam Dziekaniak e Kirinus (2004). Para as autoras, na prática, uma ontologia define termos associados aos textos que descreve, o que os mesmos significam e axiomas formais que restringem a interpretação e o uso dos termos.

O W3C coloca que as ontologias devem prover descrições para os seguintes tipos de conceitos (BREITMAN, 2005, p. 31): classes (ou coisas) nos vários domínios de interesse; relacionamentos entre essas coisas; propriedades (ou atributos) que essas coisas devem possuir.

Uma ontologia, para Martins (2002), requer o uso de um vocabulário específico para descrever uma realidade e mais um conjunto de axiomas lógicos necessários para dar semântica ao significado pretendido pelas palavras desse vocabulário. Sendo assim, duas ontologias podem referir-se à mesma semântica, embora apresentem vocabulários distintos de representação, tal como o uso de dois idiomas diferentes que descrevem uma mesma ontologia, por exemplo. A autora afirma que para a construção de uma ontologia são utilizados os seguintes objetos:

- *entidades*, que descrevem conceitos (elementos de um domínio estudado) e providenciam uma representação lógica;
- *atributos*, que descrevem as propriedades das entidades;
- *relações*, que descrevem as ligações entre objetos no modelo (entidades e atributos);
- *restrições*, condições que o projetista impõe sobre as entidades, atributos ou relações.

Martins (2002) observa, ainda, que uma ontologia possui uma hierarquia de conceitos dentro de um domínio, as descrições de cada

conceito e as propriedades definidas por atributos do tipo valor. Consiste, geralmente, de uma taxonomia² e de um conjunto de *regras de inferências*. A taxonomia define classes, subclasses e as relações entre elas, e o conjunto de regras de inferência fornece o mecanismo de manipulação dos objetos das classes utilizando raciocínio lógico.

Sendo assim, para Cunha (2002), uma ontologia fornece um entendimento comum e compartilhado de um domínio, que pode ser comunicado através de pessoas e sistemas de aplicação, tornando-se fator chave para o desenvolvimento *da Web Semântica*. O autor salienta:

A ontologia tem um papel crucial no sentido que permite o acesso, a interoperação e a comunicação baseados em conteúdo, fornecendo à *Web* um nível de serviço qualitativamente novo, que consideramos na *Web Semântica*. Ela une em rede incríveis porções do conhecimento humano, complementando-as com capacidade de processamento de máquina. (CUNHA, 2002, p. 14).

O autor supracitado ressalta, também, que as ontologias vêm sendo aplicadas no gerenciamento de dados semi-estruturados como um suporte semântico para o acesso de determinadas informações de interesses presentes em um conjunto de fontes semi-estruturadas.

Para a construção de ontologias, Freitas (2004, p. 27) observa que alguns princípios, se usados com precisão, garantem sua qualidade:

- *clareza*: os programas usam diferentes modelos e abstrações na resolução de seus problemas. Na definição do conhecimento, deve-se ter a objetividade de definir apenas o que se presume ser útil na resolução da classe de problemas a ser atingida. As definições completas, com condições necessárias e suficientes devem ter precedência sobre definições parciais;
- *legibilidade*: as definições devem corresponder com as definições correntes e informais. A ontologia deve usar um vocabulário compartilhável (geralmente o jargão e a terminologia usados por especialistas do domínio);
- *coerência*: as inferências derivadas da ontologia definida devem ser corretas e consistentes do ponto de vista formal e informal com as definições;
- *extensibilidade*: a ontologia deve permitir extensões e especializações monotonicamente e com coerência, sem a necessidade de uma *revisão de teoria*, que consiste na revisão lógica automática de uma base de conhecimento em busca de contradições;

² Uma taxonomia, segundo Daconta apud Breitman, 2005, é a classificação de informações no formato de uma hierarquia de acordo com relacionamentos estabelecidos com entidades do mundo real que representam.

- *mínima codificação*: devem se especificados conceitos genéricos independente de padrões estabelecidos para mensuração, notação e codificação, garantindo a extensibilidade. Essa genericidade é limitada pela clareza.
- *mínimo compromisso ontológico*: com a finalidade de maximizar o reuso, apenas o conhecimento essencial deve ser incluído, gerando a menor teoria possível acerca de cada conceito, e permitindo a criação de novos conceitos, mais especializados ou estendidos.

Nesse contexto, Souza e Alvarenga (2004) afirmam que as ontologias se apresentam como um modelo de relacionamento de entidades e suas interações, em algum domínio particular do conhecimento ou específico a alguma atividade. O objetivo de sua construção é a necessidade de um *vocabulário compartilhado* para se trocarem informações entre os membros de uma comunidade, sejam eles seres humanos ou agentes inteligentes.

Lembrando Breitman (2005, p. 44), as ontologias "[...] servem para estruturar e compartilhar conhecimento, não para representar inteligência." Para a autora, o papel da ontologia, na *Web Semântica*, é explicitar o vocabulário utilizado e possibilitar um padrão para o compartilhamento da informação, fornecendo um modelo comum que permita aos agentes de *software* trocar informações de modo significativo.

A propósito de Freitas (2004), as ontologias servem não só como vocabulário de comunicação entre agentes, mas também na definição e organização apropriadas de conceitos, relações e restrições. Além disso, a utilização de ontologias permite um ganho de expressividade e flexibilidade, uma vez que o conhecimento sobre uma classe não se circunscreve a termos e palavras-chave como nos mecanismos de busca, mas a qualquer fato que diga respeito às páginas, tais como estrutura, regiões e conceitos nelas contidos.

Tesauros e ontologias

A ciência da informação propõe a organização do conhecimento, e essa organização, como já dito, é fundamentada na classificação de conceitos proposta por Aristóteles. Assim, foram formadas diversas categorias para que toda e qualquer tipo de conhecimento produzido pudesse ser enquadrado em alguma dessas classes ou categorias pré-estabelecidas.

Dessa forma, para a efetiva organização do conhecimento realizada pela área, identificou-se a necessidade da adoção de uma linguagem documentária que buscasse uma padronização da linguagem natural, eliminando a polissemia característica da mesma.

Como observam Moreira *et al* (2004), sendo a organização do conhecimento uma das principais preocupações de Aristóteles, é justamente no momento da construção das linguagens documentárias que sua filosofia se faz mais presente.

Nesse sentido, os tesouros, como linguagem documentária, foram criados com o intuito de formalizar a padronização de termos da linguagem específica de uma determinada área do conhecimento humano. Assim, as possíveis ambigüidades na utilização de termos em linguagem natural são eliminadas, permitindo que o processo de indexação e recuperação de informações fosse realizado satisfatoriamente.

Os tesouros evoluíram das listas de cabeçalhos de assuntos até os tesouros baseados em conceitos. Assim, era preciso trabalhar com um vocabulário mais específico e com uma estrutura mais depurada do que a presente em cabeçalhos de assuntos. Logo, cuidou-se de melhorar, além da especificidade, a estrutura. (GOMES *apud* SOUTO, 2003).

O tesouro é uma linguagem documentária, um *vocabulário controlado* que contém termos relacionados lógica e semanticamente.

Na área de ciência da informação, os tesouros podem ser definidos sob dois aspectos: o estrutural e o funcional, segundo a Unesco *apud* Moreira *et al* (2004, p. 3). No primeiro caso, seria:

Um vocabulário controlado dinâmico de termos relacionados semântica e genericamente, cobrindo um *domínio específico do conhecimento*. Quanto à definição funcional, seria: Um dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem dos sistemas (linguagem de documentação, linguagem de informação) mais restrita.

Estipulando o controle da linguagem, a utilização de tesouros na organização do conhecimento elimina a polissemia da linguagem natural, não permitindo ambigüidades.

Um *tesouro*, segundo Breitman (2005), reúne um conjunto de relacionamentos entre termos que estão organizados em uma taxonomia. A autora define Tesouros como "[...] *uma taxonomia adicionada de um conjunto de relacionamentos semânticos (equivalência, associação, entre outros) entre seus termos.*" (BREITMAN, 2005, p. 36).

Uma taxonomia define termos, mas a única relação entre os termos é a de generalização, ou seja, os termos encontram-se, sempre, em uma hierarquia que parte de gênero para espécie, ou seja, do geral para o específico. Um tesouro é uma taxonomia na qual se acrescentam possibilidades de relacionamentos.

Quanto às relações estruturais entre os termos que o tesouro possibilita, caracterizemos as explicitadas por Gusmão (1985, p. 22):

relações de equivalência ou sinonímia. São referências cruzadas, remetem de um termo não adotado no sistema para um sinônimo ou similar que foi adotado. Para isso, usa-se a remissiva *U*, que quer dizer *USE*. Partindo-se do termo adotado, o termo não adotado é indicado pela remissiva *UF* (*used for*) ou *UP* (*usado para*). Gusmão (1985) salienta que, ao selecionar os termos preferidos (adotados) deve-se dar prioridade à forma mais compreensiva pelos usuários, aos termos científicos e/ou termos correntes:

- *relação de hierarquia*. É estabelecida para cada descritor e indica relações de superordenação ou subordinação. A de superordenação mostra os termos mais abrangentes (gerais) com eles relacionados, enquanto as relações de subordinação mostram os termos mais específicos relacionados com eles. Os símbolos utilizados são BT (*broader term*) ou TG (termo genérico), e NT (*narrow term*) ou TE (termo específico);
- *relação de associação*. É utilizada para representar termos que têm uma correlação entre si, e é representada pelos símbolos RT (*related term*) ou TR (termo relacionado). Nesse tipo de relação, os termos não representam uma hierarquização. Podemos dizer, em outras palavras, que eles têm igual importância dentro do assunto que representam.

Para Breitman (2005), um tesouro visa garantir que conceitos sejam descritos de maneira consistente, permitindo que os usuários possam refinar suas buscas e localizar a informação que necessitam. Para facilitar essa tarefa, os tesouros contam com os relacionamentos adicionais que auxiliam na organização dos dados, como explicitado anteriormente. No entanto, os tipos de relacionamentos permitidos pelos tesouros são *finitos* e pré-determinados, isto é, não são passíveis de modificações, sendo bem útil na criação de vocabulários, mas não são suficientes para modelar outros aspectos do mundo real.

Atualmente, como observam Moreira *et al* (2004), existe uma grande demanda para desenvolvimento de sistemas que trabalhem com recuperação e troca de informações e conhecimentos. Com a finalidade de servir de suporte a esses sistemas e auxiliar a suprir essa demanda, novos instrumentos para a organização de conhecimento surgem a cada dia; muitos deles, recentemente, são chamados *ontologias*.

Um dos principais objetivos do uso de ontologias no âmbito da ciência da computação é a construção de bases de conhecimento interoperáveis e melhor estruturadas. Sob a denominação de ontologias, ferramentas têm sido criadas para auxiliar a inserção e recuperação de documentos em sistemas computacionais, a extração de informações em texto de linguagem natural, em sistemas de comércio eletrônico, a troca de informações entre agentes inteligentes, a aquisição automática de conhecimento e várias outras tarefas que envolvem o uso e a representação do conhecimento (MOREIRA *et al*, 2004).

De acordo, ainda, com as autoras supracitadas, o emprego do termo ontologia para denominar uma estrutura de termos e as relações entre eles em um determinado domínio é mais comum na área da ciência da computação e, mais particularmente, na inteligência artificial. Com essa definição de ontologia, fica clara a confusão entre a definição desse termo com o conceito de tesouros, uma vez que estes, da mesma forma, podem ser considerados estruturas de termos e das relações entre eles. Contudo, Breitman (2005, p. 37), afirma:

Muitas vezes é necessário relacionar conceitos utilizando relacionamentos do tipo parte-de, membro-conjunto, fase-

processo, lugar-região, material-objeto, causa-efeito, entre muitos outros. Um tesouro não permite a seus usuários a criação destes e novos tipos de relacionamento, para tal é necessário utilizar uma ontologia. (BREITMAN, 2005, p. 37).

Sendo assim, de acordo com as afirmações de Breitman (2005), podemos argumentar que as ontologias são mais flexíveis e complexas que os tesouros porque permitem que novos relacionamentos sejam estabelecidos entre os termos, conforme necessário.

Os tesouros são linguagens pré-definidas e suas relações são finitas. Essa ferramenta é bastante útil para a padronização da linguagem de termos e relações em um domínio de conhecimento para definir termos de organização e recuperação, por exemplo, em uma base de dados.

Segundo Moreira *et al* (2004), o tesouro constitui-se em um elemento de suma importância em um sistema de recuperação de informações por cumprir o papel de: determinar quais termos podem ser usados no sistema; determinar quais termos podem ser usados na busca para que esta tenha um resultado satisfatório e, ainda, permitir a introdução de novos termos em sua estrutura de termos e relações de modo a aproximar a linguagem do usuário à linguagem do sistema, realizando alterações de sentidos dos termos existentes.

Os tesouros podem ser utilizados por indexadores para atribuir uma terminologia consistente a várias bases de dados e, ainda, pelos usuários dessas bases na definição dos termos de busca, uma vez que, pela sua estrutura de termos e suas relações auxilia a encontrar o melhor termo que representa um assunto. Entretanto, no contexto da *Web Semântica*, os tesouros parecem não ser adequados, justamente devido à sua falta de flexibilidade.

As ontologias, por sua vez, podem ser redefinidas ou adaptadas em determinados contextos, e suas relações são extensíveis, conforme necessário.

Moreira *et al* (2004) afirmam que, usando linguagens para ontologias é mais fácil registrar certas propriedades do que com o uso dos tesouros. Por outro lado, essa diferença de expressividade não é tão significativa para a tarefa de indexação ou busca de documentos.

Nesse sentido, percebemos que, se ontologias e tesouros foram criados em contextos distintos, em situações também distintas, poderá ser observado que há casos em que o uso de tesouros é mais adequado, enquanto, em outros casos faz-se necessária a aplicação das ontologias.

Enquanto os tesouros se mostram eficientes no escopo da Ciência da Informação no que condiz à indexação e recuperação de informações, no âmbito da *Web Semântica*, para descrição de recursos, a utilização de ontologias é fundamental.

De todo modo, McGuinness *apud* Breitman (2005) definiu três propriedades essenciais de uma ontologia para a *Web* que a torna mais expressiva que uma taxonomia ou um tesouro. São elas:

- estrita hierarquia de subconceitos. Toda a instância de uma classe tem de ser uma instância do nó pai (hierarquia *tipo-de*

- formal). A organização dos termos segundo o relacionamento *tipo-de* (generalização) forma a espinha dorsal da ontologia;
- interpretação livre de ambigüidades para os significados e relacionamentos. As propriedades de cada nó podem ser definidas pelos usuários. Essas propriedades podem ter valores restritos por uma gama de valores determinada também pelos usuários (restrições de valor). Ontologias mais sofisticadas podem contar com relacionamentos mais expressivos, tais com disjunção (macho x fêmea) e parte-de;
 - utilização de um vocabulário finito, porém extensível.

Baseando-nos nas palavras dessa autora, percebemos que, em sua concepção, a ontologia proposta para a *Web Semântica* não consistirá em uma simples adaptação dos tesouros para a organização do conhecimento na *Web*, proporcionando o controle de vocabulário e o *fechamento semântico* de seu conteúdo. Para essa autora, as ontologias deverão ser mais flexíveis e complexas que os tesouros.

Uma outra diferença importante a ser destacada é que, enquanto o tesouro é visto como uma ferramenta de *organização do conhecimento*, com a finalidade de facilitar a indexação e recuperação de informações, operando para a comunicação entre usuários e linguagens documentárias, a ontologia é tida como um instrumento de *representação do conhecimento* no ambiente computacional, voltada para o registro de conceitos de um domínio visando a inferência automatizada.

Estabelecendo, ainda, uma comparação entre ontologia e tesouro, Moreira *et al* (2004) observam que uma ontologia, como vista pela ciência da computação, é um sistema de conceitos, da mesma forma que os tesouros e, como tal, pertence ao nível epistemológico e não ao ontológico. A diferença em relação aos tesouros pode ocorrer em termos de linguagem, de nível de formalização e de propósitos.

Para a autora supracitada, os tesouros têm como propósito servir de instrumento de registro terminológico e para ser usado por *peçoas*, não para registro do conhecimento *para inferências computacionais*, caso das ontologias.

Corroborando essa idéia, Breitman (2005) afirma que o processo de classificação das informações contidas em uma ontologia deve levar em conta a possibilidade de automação dessa informação, e não a maneira com que os seres humanos organizam o conhecimento. Dessa forma:

Ontologias não refletem a maneira com que os seres humanos pensam nem classificam. Se estivéssemos buscando um modelo que refletisse o modo como os seres humanos organizam seu conhecimento, ontologias definitivamente não seriam o modelo mais adequado. (BREITMAN, 2005, p. 43).

Analisando a citação acima, percebemos que a autora faz uma distinção entre a função das ontologias e a clássica organização do

conhecimento possibilitada pela linguagem controlada, com o uso dos tesouros. No entanto, como dito anteriormente, parece haver uma contradição nas palavras da autora quando afirma que os metadados permitirão a adoção de um vocabulário controlado, necessário para a realização da *Web Semântica* e, ainda, na observação que as ontologias não refletem o modo formal de organização do conhecimento.

Nesse sentido, cabe notar que, se um tesouro estabelece um vocabulário controlado, ele assemelha-se às ontologias, embora ambos sejam utilizados em contextos e com princípios distintos, conforme destacou Moreira *et al* (2004, p. 22):

Os tesouros da ciência da informação e as ontologias da ciência da computação possuem *origens e propósitos* distintos. O primeiro nasceu como instrumento prático para auxiliar na indexação e busca de documentos e o segundo da necessidade de descrever os objetos digitais e suas relações. Pode-se dizer que existem alguns pontos de contato entre essas origens, uma vez que estão relacionadas com a descrição de alguma entidade: assunto de uma área no primeiro caso e objetos e relações no segundo. No entanto, as diferenças também deixaram suas marcas, influenciando na forma final do instrumento de cada vertente. Na ciência da computação a situação é um pouco mais indefinida. Aparentemente *tudo que modela um segmento da realidade pode ser denominado de ontologia*, uma vez que é a palavra da moda. Neste caso, até mesmo os tesouros podem ser enquadrados como *ontologias terminológicas*. (grifo nosso).

Como podemos perceber, a autora distingue ontologias e tesouros quanto às suas origens e propósitos, embora finalize sua assertiva dizendo que o tesouro pode ser considerado uma ontologia terminológica. E, ainda, como a autora observou, os tesouros e as ontologias foram criados em contextos e com finalidades distintas. Contudo, como ambas as ferramentas foram idealizados com o intuito de descrição e padronização acaba ocorrendo uma certa confusão na definição e conceitualização entre tesouros e ontologias.

Alguns pesquisadores apontados por Moreira *et al* (2004), tais como Jasper e Milke Uschold (1999) e Fensel (2001), consideram os tesouros como ontologias simples, uma vez que uma ontologia complexa exige uma riqueza maior de relações do que as tradicionalmente apresentadas em um tesouro.

Na ciência da informação os tesouros são, há muito tempo, amplamente utilizados como linguagem documentária. A definição e os conceitos dos tesouros estão, portanto, consolidados. Na ciência da computação as ontologias consistem em tecnologias recentes com definições e conceitos ainda não difundidos e plenamente cristalizados.

Moreira *et al* (2004) afirmam que os tesouros podem ser entendidos como sendo um tipo de ontologia voltada para a organização de termos. Da mesma forma, poderíamos entender as ontologias como tesouros extensíveis e reutilizáveis no escopo na computação. Afirmações como essas é que denotam a existência de conceitos e diferenças

imprecisos entre tesouros e ontologias. Se, por um lado, um tesouro é visto como uma ontologia simples e, por outro lado, as ontologias são entendidas como tesouros complexos e flexíveis, podemos notar que, embora consistam em diferentes ferramentas, tesouros e ontologias, como vimos, possuem características em comum que permitem a comparação de ambos na definição de um dos termos.

Dessa forma, percebemos que, algumas vezes, o termo tesouro, utilizado comumente no âmbito da ciência da informação, é utilizado como *analogia* ao significado de ontologia, uma vez que, como vimos anteriormente, essas ferramentas possuem características comuns. Entretanto, analisando a literatura foi possível notar que, embora semelhantes em alguns aspectos, tesouros e ontologias são diferentes em muitos outros, tais como em suas origens, princípios, graus de expressividade e relações, além de se aplicarem em contextos também distintos.

Conclusão

As novas tecnologias de informação e comunicação afetam significativamente todos os setores da sociedade, e a disponibilização de conhecimento e informações no ciberespaço, ambiente da *Web*, faz surgir novos paradigmas de organização e recuperação de dados.

O fato de as páginas da *Web* atual não conterem dados sobre si mesmas dificulta a recuperação de conhecimento e informações relevantes. Ao solicitar uma busca por determinada palavra, o sistema não faz distinção entre os diversos sentidos que uma mesma palavra pode representar, ou seja, a diferença de significado afeta significativamente os resultados. Assim, os mecanismos de busca frequentemente apresentam grande número de páginas encontradas, porém com pouca precisão.

Isso acontece porque as palavras indexadas pelas ferramentas de busca podem ter diversos sentidos (polissemia) em diferentes contextos. Uma vez realizada a busca, fica a cargo do usuário determinar a *semântica*, ou seja, o sentido e o significado do conteúdo obtido como resultado.

Para Breitman (2005), a grande verdade é que a Internet se desenvolveu mais rapidamente como um meio para a troca de documentos entre pessoas, em vez de um meio que fomentasse a troca de dados que pudessem ser processados automaticamente. Como consequência disso, o conteúdo semântico das páginas, ou seja, seu significado é codificado apenas pelo homem.

Falta, na *Web*, uma forma de organizar os dados de maneira que estes possam ter seu sentido identificado pela própria máquina no momento da busca.

Nesse contexto, podemos perceber que a *Web Semântica* está sendo projetada para tentar resolver um problema apresentado na *Web* atual: a pesquisa realizada pelas ferramentas de busca.

A *Web Semântica*, para Dziekaniak e Kirinus (2004), visa incorporar semântica aos dados, o que proporcionará não somente aos usuários

entenderem as informações como também às máquinas. Ela pretende fornecer estruturas e dar significado semântico ao conteúdo da *Web*, criando um ambiente onde agentes de *software* e usuários possam trabalhar de forma cooperativa.

Dessa forma, os mecanismos de busca trariam resultados adequados à semântica dos termos procurados pelo usuário, aumentando (senão otimizando) a relevância dos mesmos.

Para a efetivação dessa extensão da *Web* foram propostas diversas tecnologias, e dentre essas cabe-nos destacar a criação de *ontologias*, utilizada como um vocabulário controlado, visando atribuir sentido e significado ao conteúdo dos documentos, atuando como ferramenta de representação do conhecimento.

No decorrer da pesquisa, procurou-se verificar se as ontologias seriam utilizadas apenas para representar o assunto de uma página *Web* através de termos contextualizados ou se tentariam controlar o vocabulário da *Web* como um tesouro, que estabelece uma linguagem controlada.

Nesse sentido, foi possível perceber que as ontologias, embora sejam definidas por alguns autores como vocabulários controlados, assim como os *tesouros*, não pretendem realizar, na *Web*, o que os tesouros fizeram na tradicional organização do conhecimento: o fechamento semântico no significado.

Isso porque as ontologias são mais complexas e *flexíveis* que os *tesouros*, porque permitem que novos relacionamentos sejam estabelecidos entre os termos, conforme necessário, e são criadas ou adaptadas (quando reutilizadas) ao contexto do documento que descrevem. Os *tesouros*, ao contrário, possuem termos, conceitos e relacionamentos finitos e bem definidos.

Cabe notar que os termos ontologia e tesouros são mais utilizados no âmbito da ciência da computação e ciência da informação, respectivamente. Obviamente, com a crescente interdisciplinaridade entre as áreas do conhecimento, acabou ocorrendo um conflito de definições dessas ferramentas devido ao fato de as mesmas apresentarem algumas características em comum.

Sendo assim, percebemos que alguns autores da ciência da informação, na tentativa de definir e/ou conceituar ontologia, acabam comparando-a com um tesouro, em analogia, de forma a explicar como as ontologias pretendem atuar no contexto da *Web Semântica*.

Da mesma forma, alguns autores relacionados à ciência da computação afirmam que os tesouros são ontologias simples e que, portanto, as ontologias podem ser estendidas, tornando-se mais expressivas, de acordo com a necessidade e o contexto.

Provavelmente o termo tesouro é utilizado como analogia na definição de ontologias por tratar-se de um termo cristalizado, enquanto ontologia, fora do escopo da filosofia, é um termo recente, ainda em vias de definição.

No entanto, com a realização do trabalho ficou claro que, apesar de existirem pontos comuns entre ontologias e tesouros, trata-se de ferramentas distintas, com origens e propósitos distintos.

Enquanto os *tesouros* são utilizados para realizar o fechamento semântico no significado, eliminando a polissemia característica da linguagem natural, traduzindo os termos para uma linguagem documentária padronizada, as ontologias pretendem servir de vocabulário controlado, porém flexível, para que os conceitos descritos em páginas *Web* sejam únicos e, assim possam ser passíveis de compartilhamento entre diferentes agentes, permitindo a adição de semântica aos documentos disponíveis na rede.

Com o advento do ciberespaço, o profissional da informação dispõe de um novo meio e espaço de trabalho: a *Web*. A *Web* modificou os parâmetros de organização do conhecimento e tratamento da informação e, nesse contexto, a *Web Semântica* propõe uma nova forma de representação do conhecimento e da informação, uma nova revolução, um novo paradigma.

Sendo assim, se a *Web* já representou uma verdadeira revolução no que condiz à produção, representação e recuperação de informações e conhecimento com suas tecnologias e possibilidades, a *Web Semântica* parece uma proposta de revolucionar a *Web*.

Referências

BRANSKI, Regina Meyer. Recuperação de informações na *Web*. *Perspect. Ciênc. Inf.*, Belo Horizonte, v. 9, n. 1, p. 70-87, jan./jun. 2004.

BREITMAN, Karin. *Web Semântica: a Internet do futuro*. Rio de Janeiro: LTC, 2005.

CENDÓN, Beatriz V. A Internet. In: CAMPELLO, B. S.; CÉNDON, B. V.; KREMER, J. M. *Fontes de informação para pesquisadores e profissionais*. Belo Horizonte: Ed. UFMG, 2003. p. 275-299.

CENDÓN, Beatriz V. Ferramentas de busca na *web*. *Ci. Inf.*, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

CHATEAUBRIAND, Oswaldo. A filosofia, a linguagem e o mundo. In: BRITO, Adriano Naves de; VALE, Oto Araújo. *Filosofia, lingüística, informática: aspectos da linguagem*. Goiânia: Ed. UFG, 1998.

CUNHA, Luiz Manoel Silva. *Web Semântica: estudo preliminar*. Campinas: Embrapa Informática Agropecuária, 2002. Disponível em: <<http://www.cnptia.embrapa.br/modules/tinycontent3/content/2002/doc18.pdf>>. Acesso em 03 fev. 2006.

DZIEKANIAK, Gisele Vasconcelos; KIRINUS, Josiane Boeira. *Web Semântica*. *Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.*, Florianópolis, n. 18, p. 20-39, 2004. Disponível em <http://www.encontros-bibli.ufsc.br/Edicao_18/2_Web_Semantica.pdf>. Acesso em 31 jan. 2006.

FACHIN, Odília. *Fundamentos de metodologia*. São Paulo: Saraiva, 2003.

FREITAS, Frederico Luiz Gonçalves de. *Ontologias e a web semântica*. Universidade Católica de Santos. 2004. Disponível em <<http://www.inf.unisinos.br/~renata/cursos/topicosv/ontologias-ws.pdf>>. Acesso em 04 maio 2006.

GUIRALD, Pierre. *A semântica*. São Paulo: DIFEL, 1980.

GUSMÃO, Heloísa Rios. *Tesouros: análise e utilização*. Niterói: UFF, 1985.

LÉVY, Pierre. *As tecnologias da inteligência: o futuro do pensamento na era da informática*. São Paulo: Ed. 34, 1993.

MARTINS, Rosane Maria. *Web Semântica: uma visão geral*. Universidade Federal do Rio de Janeiro, Núcleo de Computação Eletrônica. Rio de Janeiro, nov. 2002. Disponível em: <http://www.eng.uerj.br/~rosane/survey_generico.pdf>. Acesso em: 03 fev. 2006.

MONTEIRO, Silvana Drumond. O ciberespaço e os mecanismos de busca: novas máquinas semióticas. *Ci. Inf.*, Brasília, v. 1, n. 1, p. 31-38, jan./abr. 2006.

MONTEIRO, Silvana Drumond; PICKLER, Maria Elisa Valentim. *O ciberespaço: o termo, a definição e o conceito*. (No prelo).

MOREIRA, Alexandra; ALVARENGA, Lídia; OLIVEIRA, Alcione de Paiva. O nível do conhecimento e os instrumentos de representação: tesouros e ontologias. *Datagrama zero: Revista de Ciência da Informação*, v. 5, n. 6, dez. 2004. Disponível em <www.dgzero.org/dez04/Art_01.htm>. Acesso em 10 ago. 2006.

OLIVEIRA, Rosa Maria Vivona Bertolini. Web Semântica: novo desafio para os profissionais da informação. *In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSTÁRIAS*. Disponível em <<http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/124.a.pdf>>. Acesso em 31 jan. 2006.

SOUTO, L. F. Recuperação de informações em bases de dados: usos de tesouro. *Transinformação*, Campinas, v. 15, n. 1, p. 63-72, jan./abr. 2003.

SOUZA, Renato Rocha; ALVARENGA, Lídia. A Web Semântica e suas contribuições para a ciência da informação. *Ci. Inf.*, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.