A Web Semântica e suas contribuições para a ciência da informação

Renato Rocha Souza
Doutorando em ciência da informação. Escola de Ciência da Informação. Universidade Federal de Minas Gerais
E-mail: rsouza@eci.ufmg.br

Lídia Alvarenga
Doutora em ciência da informação (UnB). Professora titular da Escola de Ciência da Informação. Universidade Federal de Minas Gerais
E-mail: lidiaalvarenga@eci.ufmg.br

Resumo

O presente artigo apresenta o processo de atualização por que passa a World Wide Web na sua transição para o que tem sido chamado de "Web Semântica". Neste sentido, busca-se identificar as tecnologias, as organizações associadas e o embasamento filosófico e conceitual subjacentes a esta nova web. O artigo também procura apresentar as imbricações existentes com a ciência da informação e as possibilidades de ampliação de escopo dos seus objetos tradicionais de pesquisa com o aporte dos novos padrões e tecnologias que estão sendo desenvolvidos no âmbito da Web Semântica.

Palavras-chave

Web Semântica; Ciência da informação; Internet; Sistemas de recuperação da informação; Hipertexto.

Web Semantics and its contributions to information science

Abstract

This article explores the updating process that is taking place in the World Wide Web in the transition to what is being called "The Semantic Web". In this sense, we try to identify the technologies, the associated organizations and institutions, the conceptualization and the philosophy that underlie this new web. The article also tries to show the interconnections between the semantic web and the field of information science, and how the semantic web technologies can broaden the traditional information science research subjects.

Keywords

Semantic Web; Information science; Internet; Systems of information retrieval; Hypertext.

INTRODUÇÃO

Surgida no início dos anos 90, a World Wide Web*, ou simplesmente *Web*, é hoje tão popular e ubíqua, que, não raro, no imaginário dos usuários, confunde-se com a própria e balzaquiana Internet – a infra-estrutura de redes, servidores e canais de comunicação que lhe dá sustentação. Se a Internet surgiu como proposta de um sistema distribuído de comunicação entre computadores para possibilitar a troca de informações na época da Guerra Fria, o projeto da *Web*, ao implantar de forma magistral o conceito de hipertexto imaginado por Ted Nelson & Douglas Engelbart (1962), buscava oferecer interfaces mais amigáveis e intuitivas para a organização e o acesso ao crescente repositório de documentos que se tornava a Internet. Entretanto, o enorme crescimento – além das expectativas – do alcance e tamanho desta rede, além da ampliação das possibilidades de utilização, fazem com que seja necessária uma nova filosofia, com suas tecnologias subjacentes, além da ampliação da infra-estrutura tecnológica de comunicação.

Para apresentar as mudanças por que está passando a *Web* na transição para este novo patamar que tem sido chamado de "Web Semântica" e avaliar alguns dos impactos deste fenômeno, convém explorar brevemente os conceitos inerentes aos sistemas de recuperação de informações, sua funcionalidade, e estabelecer algumas categorias de análise.

A *Web* e os sistemas de recuperação de informações

A dificuldade de conceitualização do que é um sistema de recuperação de informações (SRI) advém, a princípio, da ambigüidade dos conceitos de sistema e de informação em si (Araújo, 1995). Podemos adotar, entretanto, algumas definições que façam sentido no escopo do assunto tratado e, desde já, assumimos que, ao falar de sistemas de recuperação de informações, estamos falando em tecnologias para a recuperação de informações registradas em formato impresso ou digital.

* Em uma tradução literal, "Teia de Alcance Mundial".

As metodologias e tecnologias associadas à biblioteconomia e à documentação e, mais recentemente, à ciência da informação surgiram como uma resposta às necessidades causadas pelo papel cambiante que tomou o conhecimento humano e seus registros através dos tempos (Wersig, 1993). Com a explosão de documentos disponíveis, surgiram os diversos sistemas de informação que, mediante operações de indexação, armazenamento e recuperação, buscavam organizar e prover acesso à informação registrada em documentos. Com o fenômeno contemporâneo da crescente disponibilização destes documentos em formato digital, vimos surgir e ampliarem-se os sistemas informatizados de recuperação de informações.

Prover aos usuários fácil acesso aos documentos atinentes disponíveis é o objetivo dos SRIs. Para Lancaster & Warner (1993, p. 4-5), os SRIs são uma interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários, e desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle destes; distribuição e disseminação aos usuários. Esta visão é abrangente e inclui tarefas que são normalmente associadas a atores humanos. Salton & Mcgill (1983, p. 1) e, mais tarde, Baeza-Yates & Ribeiro-Neto (1999, p. 1) definem SRIs como sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação.

Há de se distinguir os sistemas de recuperação de informações dos sistemas de recuperação de dados, nos quais basta uma determinada condição a ser satisfeita para que se tenha uma resposta exata, fruto de uma busca completa e exaustiva. A recuperação de informações traz dificuldades intrínsecas ao conceito de "informação", como a dificuldade da determinação da real necessidade do usuário e seu melhor atendimento com os documentos que fazem parte do acervo do sistema (Foskett, 1997, p.5). Isto nos traz problemas para as consultas, como baixas **revocação*** e **precisão****.

Para a representação adequada de documentos, é necessário criar sistemas de indexação eficazes, de forma que a recuperação das informações neles contidas, de acordo com as necessidades dos usuários, seja a mais significativa possível. A determinação do processo de indexação é viável no momento em que os sistemas são projetados e deve funcionar continuamente, à medida que novas informações são adicionadas ao sistema.

Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a *Web* foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade. 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. A dificuldade de 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 atualmente utilizadas nas páginas *Web* focaliza os aspectos de exibição e apresentação dos dados, de forma que a informação seja pobremente descrita e pouco passível de ser consumida por máquinas e seres humanos. É neste contexto que surge a proposta da Web Semântica.

A WEB SEMÂNTICA

"A Web Semântica não é uma *Web* separada, mas uma extensão da atual. Nela a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas". Com estas palavras, Berners-Lee (2001) define os planos de seu grupo de trabalho no World Wide Web Consortium* (W3C) para operar a transformação que irá modificar a *Web* como a conhecemos hoje. "Web Semântica" (Decker *et alii*, 2000 & Berners-Lee *et alii*, 1999) é o nome genérico deste projeto, capitaneado pelo W3C, que pretende embutir inteligência e contexto nos códigos XML utilizados para confecção de páginas *Web*, de modo a melhorar a forma com que programas podem interagir com estas páginas e também possibilitar um uso mais intuitivo por parte dos usuários.

Embora "semântica" signifique "estudo do sentido das palavras", Guiraud (1975) reconhece três ordens principais de problemas semânticos:

* Razão do número de documentos atinentes recuperados sobre o total de documentos atinentes disponíveis na base de dados.

** Razão do número de documentos atinentes recuperados sobre o total de documentos recuperados.

* Consórcio de empresas, profissionais, cientistas e instituições acadêmicas que é responsável pela criação de padrões tecnológicos que regulam a World Wide Web.

1) a ordem dos problemas psicológicos, que relaciona os estados fisiológicos e psíquicos dos interlocutores nos processos de comunicação de signos;

2) a ordem dos problemas lógicos, que estabelece as relações dos signos com a realidade no processo de significação;

3) a ordem dos problemas lingüísticos, que estabelece a natureza e as funções dos vários sistemas de signos.

Guiraud confere à terceira ordem de problemas o *status* de "semântica por excelência" (1976, p.8), mas o uso da conotação "semântica" para a *Web* ampliada está ancorado na segunda definição, e se justifica se observarmos as aumentadas possibilidades de associações dos documentos a seus significados por meio dos metadados descritivos. Além disso, as ontologias construídas em consenso pelas comunidades de usuários e desenvolvedores de aplicações permitem o compartilhamento de significados comuns.

Berners-Lee (2001) imagina um mundo em que programas e dispositivos especializados e personalizados, chamados agentes, possam interagir por meio da infra-estrutura de dados da Internet trocando informações entre si, de forma a automatizar tarefas rotineiras dos usuários. O projeto da Web Semântica, em sua essência, é a criação e implantação de padrões (*standards*) tecnológicos para permitir este panorama, que não somente facilite as trocas de informações entre agentes pessoais, mas principalmente estabeleça uma língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

Para atingir tal propósito, 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 e que esta possa ser "consumida" por outros usuários humanos ou não, de maneira automática e não ambígua. Com a existência da infra-estrutura tecnológica comum da Internet, o primeiro passo para este objetivo está sendo a criação de padrões para descrição de dados e de uma linguagem que permita a construção e codificação de significados compartilhados. Para melhor entender estes padrões e linguagens, vamo-nos debruçar a seguir um pouco mais sobre estes conceitos.

SGML, HTML e XML

Um documento na *Web* é composto por uma mistura de dados e metadados. "Meta" é um prefixo de auto-referência, de forma que "metadados" sejam "dados sobre dados". Os metadados em documentos na *Web* têm a função de especificar características dos dados que descrevem, a forma com que serão utilizados, exibidos, ou mesmo seu significado em um contexto.

A linguagem ainda utilizada atualmente para a construção da maioria das páginas *Web* é o HTML, ou HyperText Markup Language (linguagem de marcação em hipertexto). A linguagem HTML é derivada do padrão SGML (Standard Generalized Markup Language), que é, na verdade, uma metalinguagem, ou seja, uma linguagem para descrever outras linguagens. O padrão SGML é baseado na idéia de que documentos contêm estrutura e outros elementos semânticos que podem ser descritos sem que se faça referência à forma com que estes elementos serão exibidos. O conjunto de todas as *tags* – marcações sintáticas que descrevem os dados e comandos para manipulação de um documento – passíveis de serem utilizadas por uma linguagem derivada do SGML é chamado de DTD, ou Document Type Definition.

A linguagem HTML é um conjunto definido de *tags*, ou um DTD específico do SGML, e foi criada tendo em mente a necessidade de construção de documentos para serem exibidos em dispositivos de computador (na *Web*), daí sua vocação para tratar do formato que os dados contidos no documento vão assumir ao serem exibidos. Um navegador ou *browser*, ao ler um documento HTML, interpreta as *tags* que este documento contém para decidir como serão exibidos os dados também contidos. Os navegadores atuais interpretam o HTML porque o DTD para definição do HTML é fixo, e é conhecido *a priori* pelo interpretador do navegador. Assim mesmo, podemos ter navegadores diferentes interpretando definições de exibição de forma particular, com resultados distintos no dispositivo de saída. A estrutura do HTML é rígida, não existindo a possibilidade de adição de novos comandos de marcação (*tags*), sem que haja uma redefinição do DTD da linguagem e conseqüente atualização dos navegadores para que interpretem estas novas *tags*. A última especificação do HTML lançada pelo W3C foi a versão 4.0, e desde então a linguagem não tem sofrido mais modificações.

A partir das limitações do HTML e das necessidades de uma linguagem que pudesse descrever o conteúdo semântico e os significados contextuais, além da estrutura e forma de exibição de documentos, foi criado o XML (eXtensible Markup Language). O XML é uma recomendação formal do W3C e, em determinados aspectos, assemelha-se ao HTML. Ambas são derivadas do SGML e contêm *tags* para descrever o conteúdo de

um documento. Mas, enquanto o HTML tem como objetivo controlar a forma com que os dados serão exibidos, o XML se concentra na descrição dos dados que o documento contém. Além disso, o XML é flexível no sentido de que podem ser acrescentadas novas *tags* à medida que forem necessárias, bastando para isso que estejam descritas em um DTD específico; ou seja, qualquer comunidade de desenvolvedores pode criar suas marcações (*tags*) específicas que sirvam aos propósitos de descrição de seus dados. Isto possibilita que os dados sejam descritos com mais significado, abrindo caminho para embutirmos semântica em documentos da World Wide Web e nas intranets. O HTML 5.0 ou XHTML é o HTML 4.0 reescrito como se fosse um DTD específico que segue o padrão XML.

Os dados contidos nos documentos XML podem ser exibidos em uma infinidade de maneiras, dependendo do dispositivo em que serão manuseados (telas de computador, celulares, PDAs etc.). Os documentos XML não contêm, em si, as diretivas para exibição dos dados, e, para cada dispositivo-destino específico, podemos realizar uma transformação do documento originalmente em XML para um documento passível de ser exibido ao usuário ou entendido e utilizado por outro dispositivo tecnológico. Esta transformação é realizada utilizando-se a linguagem XSL (eXtensible Stylesheet Language), e cada arquivo XSL contém as definições de exibição ou leitura de um ou vários dispositivos específicos (tela do computador, tela do celular, impressora, coletores de dados, outros sistemas de informação etc.), no formato que melhor convier (tabelas, gráficos, seqüência de caracteres etc.). O arquivo XML passa por uma transformação definida pelo XSL, e o resultado é um arquivo muito semelhante a um documento HTML comum. Desta forma, o trio XML, seu DTD específico e o XSL se apresentam como um conjunto de padrões que possibilitam o armazenamento, descrição significativa, intercâmbio e exibição dos dados de forma personalizada.

O padrão XML é aceito como o padrão emergente para troca de dados na *Web*. Mas, apesar de possibilitar aos autores a criação de suas próprias *tags*, em uma perspectiva computacional, há muito pouca diferença entre as *tags* <AUTHOR> e <CREATOR>. Para que as marcações semânticas criadas sejam utilizadas de forma não-ambígua por comunidades maiores, são necessários alguns padrões de compartilhamento mais universais. O W3C e as comunidades de usuários têm procurado prover estes padrões, como abordamos em seguida.

Muitas empresas estão migrando seus bancos de dados e bases de documentos para padrões compatíveis com XML e SGML, de forma a possibilitar a interoperabilidade dos sistemas internos da companhia.

Metadados e o Dublin Core

Não basta possuir uma linguagem flexível como o XML para construir metadados. Para compartilhar um significado, é necessário que este seja consensual e inteligível de forma não ambígua entre todos os participantes de uma comunidade. Para resolver o problema da explosão de nomenclaturas diferentes e as várias situações em que a interpretação dos dados de maneira unívoca não é possível, foram criados, no escopo do projeto da Web Semântica, alguns padrões de metadados, de construção de código XML e uma nova significação para o termo ontologias, como vemos a seguir.

O padrão Dublin Core é uma iniciativa para criação de um vocabulário controlado, mesmo que limitado, para uso na *Web*, baseado no pressuposto de que a busca por recursos de informação deve ser independente do meio em que estão armazenadas. É composto de 15 elementos de metadados (DCMI, 2003) e se baseia no padrão MARC* (2003). Seus elementos são *title* (o nome dado ao recurso, ou título), *creator* (a pessoa ou organização responsável pelo conteúdo), *subject* (o assunto, ou tópico coberto pelo documento), *description* (descrição do conteúdo), *publisher* (o responsável por tornar o recurso ou documento disponível), *contributors* (aqueles que contribuíram para o conteúdo), *date* (data em que o recurso foi tornado disponível), *type* (uma categoria preestabelecida para o conteúdo), *format* (o formato no qual o recurso se apresenta), *identifier* (identificador numérico para o conteúdo, tal como uma URL**), *source* (fonte de onde foi originado o conteúdo), *language* (a linguagem em que está escrito), *relation* (como o conteúdo se relaciona com outros recursos, como, por exemplo, se é um capítulo em um livro), *coverage* (onde o recurso está fisicamente localizado) e *rights* (um ponteiro ou *link* para uma nota de *copyright*). O Dublin Core Metadata Initiative (DCMI) teve início em 1995, ganhando o nome da localidade onde se deu o encontro inicial, Dublin, Ohio, USA. Sua aceitação foi rápida e é hoje

* O MARC é um padrão para comunicação de informações bibliográficas de forma que possibilite o entendimento por dispositivos eletrônicos. É uma iniciativa da biblioteca do Congresso dos EUA.

** A URL, ou Uniform Resource Locator, é um caso particular dos URI (*Uniform Resource Identifier*), que são os endereços que identificam um "ponto de conteúdo" da World Wide Web, seja este uma página de texto, vídeo, imagem, som etc. O tipo mais comum de URI é a URL, que descreve o endereço de uma página na *Web* (o servidor que a hospeda e o nome do documento neste servidor) e o mecanismo (protocolo) utilizado para o acesso (HTTP, FTP etc.).

um padrão internacional, com participantes de mais de 20 países. Existem duas formas para o padrão Dublin Core, a forma simples e a qualificada. Enquanto Simples apenas especifica os padrões para os 15 possíveis pares de atributo e valor, a qualificada aumenta a especificidade dos metadados com informações sobre codificação e outras orientações para o processamento dos documentos.

O padrão RDF

O RDF ou *Resource Description Framework* é uma recomendação do W3C que deve vir a ser implementada na confecção de páginas da *Web* Semântica. O RDF encerra um padrão de ontologias, para a descrição de qualquer tipo de recurso Internet, como um *site Web* e seu conteúdo. O RDF estabelece na verdade um padrão de metadados para ser embutido na codificação XML, e sua implementação é exemplificada pelo RDF Schema, ou RDFS, que faz parte da especificação do padrão. A idéia do RDF é a descrição dos dados e dos metadados por meio de um esquema de "triplas" de recurso-propriedade-valor, e uma forma coerente de acesso aos padrões de metadados (*namespaces**) publicados na *Web* (como o Dublin Core, ou outro *namespace* compartilhado). Vejamos no quadro 1 um exemplo de código XML que utiliza três diferentes *namespaces*.

QUADRO 1
Exemplo de código XML

```
<?xml version="1.0"?>
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.oclc.org/DC#"
     xmlns:v="http://www.w3.org/2001/vcard-rdf/3.0#">
<rdf:Description about=" http://www.ucla.edu/~einstein"/>
   <dc:Creator>
        <rdf:Description about=" http://www.ucla.edu/staff/einstein"/>
          <v:Name> Isaac Einstein</v:Name>
          <v:Email="einstein@ucla.edu"/>
          <v:Orgname>UCLA</v:Orgname>
          <v:Orgunit>Department of Physics</v:Orgunit>
   </dc:Creator>
</rdf:Description>
</rdf:RDF>
```

Nas segunda, terceira e quarta linhas de código, vemos a referência aos *namespaces* utilizados pelo documento XML – o *namespace* do padrão RDF, o do padrão Dublin Core e o *namespace* de especificação de Vcards (Visit Cards), que padroniza a descrição dos dados comumente encontrados em um cartão de visita. Uma vez especificado um *namespace*, podemos utilizar seus descritores de forma não-ambígua ao longo do documento, fazendo sempre referência a qual deles estamos utilizando (ex: <v:Name>, <dc:Creator> ou </rdf:Description>). Podemos ter centenas ou milhares de *namespaces* de uso geral (como o da especificação Dublin Core) ou específicos (como o do padrão Vcard) publicados na *Web*, de forma que os metadados estejam sempre disponíveis, e, sempre que precisarmos de um vocabulário controlado para descrever algum domínio do conhecimento, possamos recorrer aos metadados consensuais e compartilhados. O padrão RDF, as ontologias e os *namespaces* compartilhados vão permitir que qualquer indivíduo ou organização publique informações em *sites Web* de forma que produtos de *software* ou agentes possam interpretar a informação marcada semanticamente e agir sobre esta informação de forma mais inteligente.

Em resumo, são estes alguns dos benefícios do padrão RDF:

– prover um ambiente consistente para a publicação e utilização de metadados na *web* utilizando a infra-estrutura do XML;

– prover uma sintaxe padronizada para a descrição dos recursos e propriedades dos documentos na *Web*;

– permitir que aplicações possam agir de forma inteligente e automatizada sobre as informações publicadas na *Web*, uma vez que seus significados são mais facilmente inteligíveis.

O padrão RDF ainda está em evolução, e se estudam soluções para que a descrição dos *namespaces* seja feita de forma mais inteligente e não repetitiva no escopo de um documento e, além disso, possam compreender mais propriedades. Neste âmbito, vamos falar um pouco mais sobre um tipo mais genérico de *namespace*, que são as ontologias.

Ontologias

A palavra "ontologia" deriva do grego *onto* (ser) e *logia* (discurso escrito ou falado). Na filosofia, a ontologia é uma teoria sobre a natureza da existência, de que tipos de "coisas" existem; a ontologia como disciplina filosófica estuda tais teorias. Pesquisadores da *Web* e de

* Um *namespace* (NS) define um vocabulário controlado que identifica um conjunto de conceitos de forma única para que não haja ambigüidade na sua interpretação. Os *namespaces* XML são conjuntos de tipos de elementos e atributos possíveis para cada tipo. As triplas do RDF se baseiam em *namespaces* de forma que a cada recurso seja associado uma dupla de propriedade e valor. Os *namespaces* podem ser referenciados por meio de uma URI, que se constitui em um repositório compartilhado, e não-ambíguo, onde usuários e programas de validação de código XML podem consultar a sintaxe e propriedades semânticas dos conceitos cobertos.

inteligência artificial adaptaram o termo aos seus próprios jargões, e, para eles, uma ontologia é um documento ou arquivo que define formalmente as relações entre termos e conceitos. Neste sentido, uma ontologia mantém semelhanças com os tesauros, utilizados para definição de vocabulários controlados. Nas palavras do SemanticWeb.org,

"uma ontologia é uma especificação de uma conceituação. É designada com o propósito de habilitar o compartilhamento e reuso de conhecimentos, de forma a criar 'compromissos ontológicos', ou definições necessárias à criação de um vocabulário comum".

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 humanos ou agentes inteligentes. Diversos padrões e linguagens para construção e compartilhamento de ontologias na *Web* estão sendo criados, todos baseados no XML, com algumas diferenças de sintaxe de marcação (*tags*). Alguns exemplos são o SHOE*, a Ontology Exchange Language (XOL)**, a Ontology Markup Language (OML e CKML***) e a Resource Description Framework Schema Language (RDFS)****. Existe uma proposta de extensão do RDF e o RDFS chamada OIL (Ontology Interchange Language)***** e seu sucessor DAML+OIL******.

O DAML+OIL (DARPA Agent Markup Language – Ontology Interchange Language) é uma linguagem baseada no XML, desenhada para possuir muito mais capacidade que este na descrição de objetos e no seu relacionamento; para expressar semântica e criar um alto grau de interoperabilidade entre *sites Web*. O OWL é uma linguagem de marcação semântica para publicação e compartilhamento de ontologias na *Web* e do DAML+OIL. Um exemplo de um editor que suporta a criação cooperativa de ontologias baseado na *Web* é o Webonto*******.

* http://www.cs.umd.edu/projects/plus/SHOE/

** http://ecocyc.panbio.com/xol/xol.html

*** http://www.ontologos.org/

**** http://www.w3.org/TR/PR-rdf-schema/

***** http://www.ontoknowledge.org/oil/

****** http://www.daml.org/

******* http://webonto.open.ac.uk/

Agentes

O grande poder da Web Semântica só vai se realizar quando forem criadas peças de programa que coletem conteúdo da *Web* de diversas fontes, processem estas informações e compartilhem os resultados com outros programas. Estes programas são os agentes. Embora não haja uma definição universal para o termo "agente" no âmbito da computação, podemos considerar o conceito disseminado de agentes como assistentes de tarefa, ou seja, entidades de *software* que empregam técnicas de inteligência artificial com o objetivo de auxiliar o usuário na realização de uma determinada tarefa, agindo de forma autônoma e utilizando a metáfora de um assistente pessoal.

A tecnologia de agentes permite que se repense a natureza da interação entre homem e computador, na qual esse último torna-se um parceiro do usuário, cooperando para o alcance dos objetivos traçados. Podemos esperar que o futuro da computação seja caracterizado por uma completa delegação de tarefas por parte dos usuários aos computadores, sem a necessidade de qualquer tipo de manipulação direta. A utilização de agentes possibilita a implementação de um estilo complementar de interação, chamado gerência indireta, no qual o computador se torna uma entidade ativa, dotada de certo grau de autonomia e capaz de realizar tarefas que auxiliem o usuário no desempenho de suas atividades, de acordo com seus interesses.

Em Wooldridge & Jennings (1995), é apresenta-se um conjunto de propriedades desejáveis a um agente, a saber:

– ***autonomia***, de modo a agir sem qualquer tipo de intervenção, possuindo controle sobre suas ações e estado interno;

– ***sociabilidade***, de modo a interagir com outros agentes (artificiais ou humanos) por meio de algum tipo de linguagem de comunicação;

– ***reatividade***, de modo a perceber alterações em seu ambiente, reagindo a tempo;

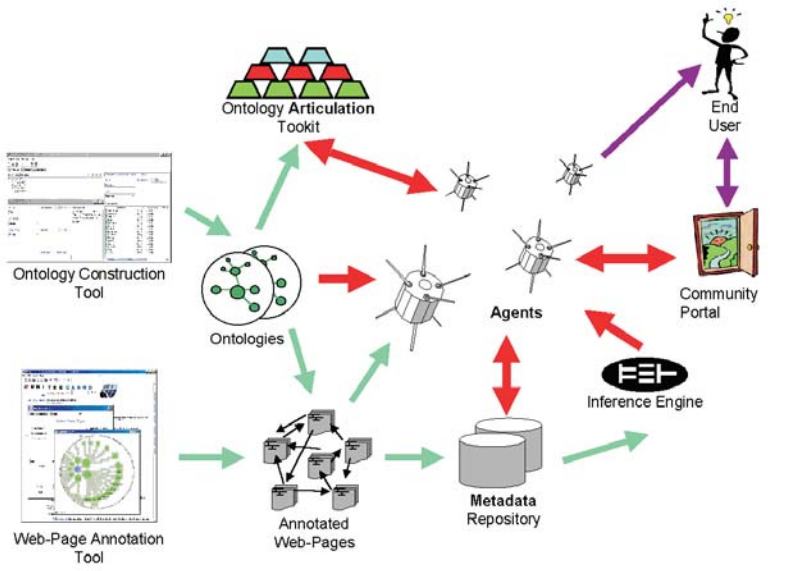
– ***proatividade***, de modo a estar apto a tomar iniciativas, em vez de simplesmente atuar em resposta ao ambiente;

– ***continuidade temporal***, ou seja, está sendo executado continuamente, ativamente ou em *background*, possivelmente captando informações sobre o usuário e sobre o ambiente, para melhor desempenhar suas funções;

– ***orientação para objetivos***, por ser capaz de interagir e desempenhar uma série diversa de ações isoladas, com objetivo de executar uma tarefa mais complexa.

Em Nwana (1996), é apresentada uma tipologia para agentes na qual estes são analisados segundo várias dimensões: mobilidade; presença de um modelo de raciocínio simbólico; exibição de um conjunto ideal e primário de atributos, tais como autonomia, cooperação e aprendizagem; papéis desempenhados pelos agentes; filosofias híbridas, decorrentes da combinação das características anteriores; atributos secundários, tais como versatilidade, benevolência, confiabilidade, qualidades emocionais, entre outros. Com base nessas características, Nwana classifica os agentes como colaborativos, móveis, de informação/Internet, reativos, híbridos, inteligentes e de interface.

FIGURA 1
O *roadmap* da Web Semântica (SemanticWeb. Org, 2001)



A efetividade destes agentes de *software* vai aumentar exponencialmente à medida que mais conteúdo marcado semanticamente e passível de ser "entendido" por máquinas estiver disponível. A Web Semântica promete esta sinergia: mesmo os agentes que não tenham sido expressamente desenhados para trabalhar em conjunto poderão trocar informações entre si, quando houver semântica embutida nestes dados.

A WEB SEMÂNTICA: A *WEB* SE APROXIMA DE UM GRANDE SRI

A partir dos conceitos de sistemas de recuperação de informações e das tecnologias apresentadas, vamos entender um pouco mais o grande panorama da Web Semântica, com a ilustração a seguir (figura 1).

Na figura 1, que ilustra o *roadmap* da Web Semântica (SemanticWeb.Org, 2001), podemos entender como as tecnologias se articulam entre si e como a Web Semântica aproxima a *Web* da funcionalidade plena de um sistema de recuperação de informações. Vamos associar as várias entidades representadas e suas funcionalidades associadas a seguir.

No âmbito da representação e indexação dos documentos, temos as **ferramentas e tecnologias para anotação semântica das páginas *web*** (Web-Page annotation Tools) e para construção de **ontologias** compartilhadas (Ontology Construction Tools). Estas ferramentas possibilitarão a existência cada vez mais ampla e disseminada de **páginas *web* marcadas semanticamente** (Annotated Web-Pages) com **metadados** descritos em **namespaces** de domínio público (Metadata Repository) e com conteúdo semântico compartilhado em seu significado pelas comunidades e usuários da *web* através das **ontologias**. As ontologias criadas serão articuladas entre si por meio de ferramentas específicas e **meta-ontologias** (Ontologies Articulation Toolkits). Com uma estratégia padronizada de indexação, podemos projetar sistemas mais funcionais para recuperação da informação armazenada.

No âmbito da recuperação e uso dos documentos, os **agentes**, associados aos **mecanismos de busca e inferência** (Inference Engine) executarão o *harvesting* (colheita) de informações nos documentos anotados semanticamente de maneira eficaz, porque serão capazes de "compreender" seus conteúdos, de modo que a informação seja mais significativamente utilizada pelos **usuários** (humanos e não humanos) da *Web*. Estes poderão acessar estas novas tecnologias por meio dos **portais comunitários** (community portals) ou mesmo dos portais corporativos das organizações. Podemos esperar que a *Web* tenha grande melhoria dos índices de revocação e precisão no atendimento das necessidades de informação, porque a semântica embutida nos documentos permitirá aos dispositivos de recuperação evitar os problemas comuns de polissemia e sinonímia,

além de considerar as informações em seus contextos de significado.

A infra-estrutura da Internet e as intranets, no âmbito das várias organizações, serão os dispositivos responsáveis pelo armazenamento e os canais por onde ocorre a disseminação dos documentos, neste grande sistema de informação. As tecnologias para implementação, assim como os protótipos destas ferramentas, já se encontram disponíveis, e o processo de atualização da *Web* está em pleno curso. A Web Semântica não trata de uma revolução, mas sim de uma evolução da *Web* como a conhecemos hoje. Trata-se principalmente da adoção de padrões de metadados e de compartilhamento destes padrões, de forma que possamos melhor utilizar o vasto repositório de informações disponível da *Web* de maneira mais produtiva, ágil e significativa.

A WEB SEMÂNTICA E A CIÊNCIA DA INFORMAÇÃO

Como dissemos, acreditamos que, na convergência destas tecnologias e ferramentas apresentadas, podem surgir alternativas para suportar um uso mais significativo e eficaz do grande patrimônio disponível nas redes de informação. Mas o que não foi dito é que estes padrões que estão sendo criados não limitam seu escopo de aplicabilidade à *Web*. A palavra de ordem que se iniciou com o XML é "interoperabilidade", ou seja, a possibilidade de sistemas diferentes "conversarem" entre si. Tudo indica que os padrões que estão sendo desenhados para esta nova *Web* também sejam adotados na arquitetura de bibliotecas digitais e de novos sistemas de informação. Podemos esperar que estas tecnologias também tragam mudanças para a área e a atividade dos profissionais da ciência da informação. Podemos exemplificar algumas atividades bem específicas que serão possibilitadas ou melhoradas com esta nova *Web*:

Projetos de novos e melhorados motores de busca

Com a marcação semântica das informações, há uma nova miríade de possibilidades para o projeto de mecanismos de recuperação de informações. Nota-se uma preocupação atual por parte dos maiores motores de busca em se preparar para essa outra versão da *Web* (Marchiori, 1998), que é constituída gradualmente de um número cada vez maior de documentos marcados semanticamente. Faz parte do escopo da ciência da informação o estudo de processos de indexação e recuperação de informações e, nesta perspectiva, é bem provável que venhamos a confrontar nossas linguagens artificiais de indexação com as metodologias de marcação semântica dos dados representadas pelos metadados e *namespaces* da Web Semântica e, também, da lógica formalizada do XML e do RDF.

Construção de novas interfaces com o usuário para sistemas de informação

O estudo de interfaces dos sistemas de informação, como os motores de busca, com o usuário, ganha um novo impulso com aumentadas possibilidades da Web Semântica, uma vez que a lógica intuitiva e natural do RDF permite que projetemos interfaces para sistemas de informação de forma mais intuitiva e coerente com o funcionamento cognitivo dos seres humanos. Além disso, com os agentes inteligentes, poderemos aprimorar e personalizar a utilização dos perfis de usuários para que a interação destes com os sistemas seja mais significativa e ágil. A lógica de triplas do RDF casa-se sobremaneira com a construção de mapas conceituais (Novak, 1977), e podemos adotar estratégias de visualização como as geometrias hiperbólicas (Lamping, 1995).

Construção automática de tesauros e vocabulários controlados

Devido à semântica genérica e formalizada do RDF (Ora Lassila *et alii*, 1999), às possibilidades de se embutir significado nos documentos e à disponibilização de ontologias em diversas áreas do conhecimento, podemos esperar o surgimento de novas metodologias automatizadas para criação de tesauros e vocabulários controlados, a partir da análise das marcações semânticas dos documentos e das relações tríplices de **recurso**, **propriedade** e **valor,** explicitadas pelo RDF.

Indexação automática de documentos

Por meio das ontologias e dos metadados utilizados, compartilhados e validados entre comunidades de interesse, podemos engendrar novas metodologias para analisar automaticamente a atinência de documentos e assim classificá-los de maneira automática ou semi-automática.

Gestão do conhecimento organizacional

De acordo com Teixeira Filho (2000), a gestão do conhecimento organizacional nasce da confluência entre tecnologia da informação e administração e se posiciona entre os campos da cultura organizacional, estratégia empresarial e sistemas de informação de uma organização. Outros autores poderiam acrescentar o campo da educação corporativa e de recursos humanos, e é um dos campos de estudo da ciência da informação. Podemos apontar a grande confluência das tecnologias tornadas disponíveis pela Web Semântica e as

necessidades de gestão do conhecimento organizacional. Com o aumento das possibilidades de recuperação de documentos e da interoperabilidade entre os sistemas, podemos esperar maior funcionalidade de portais corporativos, tecnologia-símbolo da gestão do conhecimento. Com as ontologias comunitárias e da padronização dos metadados, torna-se mais fácil a tarefa de explicitar, classificar e armazenar o conhecimento produzido pelos ativos de capital intelectual da organização.

Gestão da Informação Estratégica e da Inteligência Competitiva

Segundo Cronin (1990), as atividades de gestão de recursos de informação são vitais para acompanhamento dos ambientes externo e interno das organizações e, conseqüentemente, para a gestão estratégica do negócio. Dentre estas atividades, podemos citar a análise contínua de informações sobre indicadores selecionados publicada nas redes de comunicação como a Internet. A tecnologia dos agentes promete automatizar e agilizar a colheita destas informações, por meio da análise de dados que alimentarão *data marts* e *data warehouses**, que, por sua vez, constituirão uma fonte de informações para auxílio na tomada de decisão.

CONCLUSÕES

O objetivo deste artigo é, além de oferecer uma amostra das tecnologias e inovações que surgem com a Web Semântica, apontar as confluências entre o campo da ciência da informação, com sua tipologia e teoria sobre os sistemas de recuperação de informação, e a filosofia e as tecnologias que estão embutidas no projeto desta nova e atualizada *Web*. O estudo das possibilidades que se abrem e a compreensão de que todo o embasamento filosófico, metodológico e conceitual da Web Semântica parte do núcleo duro da ciência da informação nos impelem a demarcar e arrebanhar os legítimos territórios do saber e a buscar ativamente uma atuação no desenho destes novos panoramas informacionais. Precisamos hoje trilhar um caminho rumo a uma nova e necessária valorização da área de ciência da informação, que oferece teoria, metodologias e competências que compõem a quintessência daquilo que se espera dos trabalhadores e pesquisadores de uma sociedade baseada em informação e conhecimento. E a importância da *Web* e das demais redes digitais de troca de informações no panorama mundial são amostras de como a atividade de organização da informação é necessária para a evolução dos indivíduos, organizações e da sociedade em geral.

Artigo recebido em 23-10-2003 e aceito para publicação em 24-04-2004.

* Um *data warehouse* é uma coleção de dados não-volátil, crescente no tempo, integrada e orientada ao negócio, para dar suporte a decisões gerenciais (Inmon, 1996). O *data mart* segue a mesma filosofia, mas tem abrangência menor.

REFERÊNCIAS

ARAÚJO, Vânia M.R.H. Sistemas de recuperação da informação: nova abordagem teórico conceitual. *Ciência da Informação*, Brasília, v. 24, n. 1, 1995. Disponível em: < > Acesso em: 07 fev. 2003.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval.* New York : ACM, 1999. 511 p.

BERNERS-LEE, T. *et al*. *The semantic toolbox*: building semantics on top of XML-RDF. Disponível em: <http://www.w3.org/DesignIssues/Toolbox.html>. Acesso em: jun. 2003.

BERNERS-LEE, T., LASSILA, Ora; HENDLER, James. *The semantic web*. *Scientific America*, Maio 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: jun. 2003.

CRONIN, Blaise. Esquemas conceituais e estratégicos para a gerência da informação. *Revista da Escola da Biblioteconomia da UFMG*, Belo Horizonte, v. 19, n. 2, p. 195-220. 1990.

DECKER, S. *et al.* The semantic web: the roles of XML and RDF. *IEEE Expert, v.* 15, n. 3. Oct. 2000.

DUBLIN CORE METADATA INITIATIVE. Disponível em: <http://dublincore.org.>. Acesso em: jun. 2003.

ENGELBART, Douglas. *Augmenting human intellect*: a conceptual framework. Disponível em: < http://www.liquidinformation.org/engelbart/62_paper_full.pdf >. Acesso em: ago. 2003.

FOSKETT, A. C. *The subject approach to information*. 5. ed. London : Library Association, 1997. 119 p.

GUIRAUD, Pierre. *A semântica*. 2. ed. Rio de Janeiro : Difel, 1975. 133 p.

HERMANS, B. *Inteligent software agents on the Internet*: an inventory of currently offered functionality in the information society & a prediction of (near-) future developments, Tilburg, Holanda : Tilburg University, 1996. Disponível em: <http://www.hermans.org/agents>. Acesso em: jun. 2003.

INMON, Willian. *Building the data warehouse*. 2. ed. New York : John Wiley, 1996. 401 p.

LAMPING, J; RAO, R.; PIROLLI, P. *A Focus+context technique based on hyperbolic geometry for visualizing large hierarchies*. Disponível em: <http://www.acm.org/sigchi/chi95/proceedings/papers/jl_bdy.htm>. Acesso em: jul. 2003.

LANCASTER, F. W.; WARNER, A. J. *Information retrieval today*. Information Resources, 1993.

LASSILA, Ora; SWICK Ralph R. *Resource description framework (RDF) model and syntax specification*: recommendation W3C, Feb. 1999. Disponível em: <http://www.w3.org/TR/1999/RECrdf-syntax-19990222>. Acesso em: jun. 2003.

LAWRENCE, Steve. Context in web search. *IEEE Data Engineering Bulletin*, v. 23, n. 3, p. 25-32, 2000. Disponível em: <http://citeseer.nj.nec.com/lawrence00context.html>. Acesso em: jun. 2003.

MARC standards. Disponível em: <http://www.loc.gov/marc/>. Acesso em: jun. 2003.

MARCHIORI, Massimo. *The limits of web metadata, and beyond*. *In*: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 7., 1998. Computer networks and ISDN systems. **Proceedings...** [S. l. : s. n.], 1998. v. 30. p. 1-9.

NELSON, T. H. *Literary machines*. Sausalito, CA : Mindful, 1982.

NOVAK, J. D. *A theory of education*. Ithaca, NY : Cornell University, 1977.

NWANA, H.; *Software agents*: an overview. *Knowledge Engineering Review*, v. 11, n. 3. p. 205-244, 1996.

RAGHAVAN, P. *et al*. *Finding anything in the billion page web*: are algorithms the key? Toronto : WWW8, 1999.

SALTON, Gerard; MCGILL, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book, 1983. 448 p.

SEMANTIC web. Disponível em: <http://www.semanticweb.org/about.html>. Acesso em: jun. 2003.

TEIXEIRA FILHO, J. *Gerenciando conhecimento*. Rio de Janeiro : Senac, 2000.

WEB architecture: *describing and exchanging data*. Disponível em: <http://www.w3.org/1999/04/WebData>. Acesso em: jun. 2003.

WERSIG, Gernot. Information science: the study of postmodern knowledge usage. *Information Processing & Management*, Oxford, U.K. v. 29, p. 229-239, Mar. 1993.

WHAT is computer terminology. Disponível em: <http://www.whatis.com>. Acesso em: jun. 2003.

WOOLDRIDGE, M.; JENNINGS, N. Intelligent agents: theory and practice. *Knowledge Engineering Review,* v. 10, n. 2, p. 115-152, 1995.

WOOLDRIDGE, M; JENNINGS, N. (Ed.). *Agent technology*: foundations, applications, and markets. Berlim : Springer-Verlag, 1998.