WEB SEMÂNTICA
SEMANTIC WEB

Gisele Vasconcelos Dziekaniak, Mestre
Professora Substituta do Departamento de Biblioteconomia e História
Fundação Universidade do Rio Grande
Josiane Boeira Kirinus, Mestre
Professora do Departamento de Ciências da Economia e Informática
URCAMP, São Borja, RS.

RESUMO
O trabalho aborda a Web Semântica: a nova versão da web que está em desenvolvimento, através de projetos como o Scorpion[1] e o Desire[2]. Estes projetos buscam organizar o conhecimento armazenado em seus arquivos e páginas web, prometendo a compreensão da linguagem humana pelas máquinas na recuperação da informação, sem que o usuário precise dominar refinadas estratégias de buscas. O artigo apresenta o padrão de metadados Dublin Core como o padrão mais utilizado atualmente pelas comunidades desenvolvedoras de projetos na área da Web Semântica e aborda o RDF como estrutura indicada pelos visionários desta nova web para desenvolver esquemas semânticos na representação da informação disponibilizada via rede, bem como o XML enquanto linguagem de marcação de dados estruturados. Revela a necessidade de melhorias na organização da informação no cenário brasileiro de indexação eletrônica a fim de que o mesmo possa acompanhar o novo paradigma da recuperação da informação e organização do conhecimento.

PALAVRAS-CHAVE: Web Semântica. Organização do Conhecimento. Metadados. Dublin Core. RDF. XML. Projeto Scorpion. Projeto Indexa.

1 INTRODUÇÃO

A Web Semântica surge como uma possível solução para a estruturação semântica dos dados na web, viabilizando o processamento da informação por parte das máquinas. Berners-Lee (2001), idealizador da nova web, cita um exemplo do que a Web Semântica será capaz de fazer. Neste exemplo, um usuário pede ao computador que ele encontre um médico de uma determinada área da medicina e dá à máquina algumas restrições, tais como: este médico deve ter seu consultório no mesmo bairro que o usuário e deve estar ligado à comunidade acadêmica. O computador navega pela rede e encontra algumas opções. De uma maneira inteligente e automática, ele deve comparar a agenda do usuário com a agenda do médico e oferecer opções de horários para consulta. O usuário só terá o trabalho de escolher o horário que melhor lhe convém.

[1] Página projeto Scorpion: < orc.rsch.oclc.org:6109/>

A Web Semântica representa a evolução da web atual. Enquanto a web tradicional 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 são capazes de operar eficientemente sobre as informações, podendo entender seus significados. Desta maneira, elas irão auxiliar os usuários em operações na web.

A Web Semântica visa incorporar semântica às informações. Isso proporcionará não somente aos usuários entenderem as informações como também as máquinas. Ela pretende fornecer estruturas e dar significado semântico ao conteúdo das páginas web, criando um ambiente onde agentes de software e usuários possam trabalhar de forma cooperativa.

Neste novo contexto, a web será capaz de representar associações entre coisas que em princípio poderiam não estar relacionadas. Segundo Berners-Lee (2001) 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.

Estas regras são especificadas através de ontologias, que permitem representar explicitamente a semântica dos dados. Através dessas ontologias é possível elaborar uma rede enorme de conhecimento humano, complementando o processamento da máquina e melhorando qualitativamente o nível de serviços na web.

Segundo Hendler (2001), a Web Semântica pode ser considerada como a composição de um grande número de pequenos componentes ontológicos que apontam entre si. Dessa forma, companhias, universidades, agências governamentais e grupos de interesses específicos procurarão ter seus recursos web ligados a um conteúdo ontológico, já que ferramentas poderosas serão disponibilizadas para intercambiar e processar essas informações entre aplicações web.

Entretanto, é preciso ressaltar que a semântica não está somente relacionada ao conteúdo de um recurso, mas também à forma de como este se relaciona com os demais recursos na web. Portanto, é essencial que os recursos disponibilizados sejam expressivos o bastante para que as máquinas ou agentes sejam capazes de processar e entender o real significado do dado, intermediando as necessidades de cada usuário e as fontes de informações disponíveis.

---

[2] Página projeto Desire: <http://www.ariadne.ac.uk/issue5/desire/>

O objetivo da Web Semântica é estruturar o conteúdo que está solto na Internet. Para isto é necessário que agentes[1] percorram a rede, página a página para executar tarefas consideradas sofisticadas para o usuário. Esses agentes serão capazes de identificar o significado exato de uma palavra e as relações lógicas entre várias palavras.

Para os computadores entenderem o conteúdo da web é necessário que eles consigam ler dados estruturados[2] e tenham acesso a conjuntos de regras que o ajudem a conduzir seus raciocínios. As páginas web terão de ser escritas numa linguagem nova e serem entendidas por diferentes sistemas.

Algumas tecnologias foram desenvolvidas para a Web Semântica, tais como o XML, linguagem de marcação que permite aos usuários criarem *tags*[3] personalizadas sobre o documento criado, diferentemente do HTML, que possui estrutura de *tags* fixas, impedindo a criação de novos tipos de descritores (vide seção 3.2 em que se aborda o XML de forma mais específica).

Outra tecnologia utilizada pela Web Semântica é o RDF, que trabalha com um trio de informação o qual expressa o significado das informações. Cada componente do trio tem sua própria finalidade, em analogia ao sujeito, verbo e objeto de uma frase e recebe uma identificação URI[4] (vide seção 3.1 deste artigo que volta a abordar o RDF, com maior aprofundamento).

Na linguagem humana uma palavra pode assumir vários significados, o que pode causar confusão nos sistemas. A solução é usar URIs diferentes para cada conceito. Quando acontecer de dois bancos de dados usarem URIs diferentes para um conceito é necessário que o software que vai vasculhá-los saiba quando está tratando do mesmo conceito. Neste caso são usadas as ontologias, que fornecerão o vocabulário necessário para a comunicação entre os agentes e as páginas e mostrarão as relações entre os conceitos. Na prática, uma ontologia define termos (nome de entidades num determinado contexto) associados a textos que descrevem o que os mesmos significam e axiomas formais que restringem a interpretação e o uso dos termos.

Segundo Faria e Girardi (2002), um exemplo de aplicação da Web Semântica é a Gestão do Conhecimento, área que se concentra na obtenção, manutenção e acesso ao conhecimento de uma organização, com o objetivo de aumentar a produtividade. Com o impacto da Internet e da globalização muitas empresas se inseriram no mundo virtual.

[1] Programas que percorrem a web em busca de determinada informação.
[2] Informações colocadas em campos fixos de um arquivo.
[3] Marcação de dados inserida em arquivos.

Surgiram muitos sistemas de controle de documentos on-line. Segundo os autores supracitados, esses sistemas têm fraquezas como:

- A pesquisa da informação: essas pesquisas são feitas baseadas em palavras-chave, os sistemas recuperam muita informação irrelevante devido ao uso de certas palavras em diferentes contextos;
- A extração da informação: os sistemas são requeridos para extrair informação relevante das fontes de informações, mas na realidade não o fazem;
- Manter fontes de texto fracamente estruturadas: é uma atividade difícil e que consome muito tempo, quando as fontes são grandes. Manter tais coleções consistentes, corretas, requer uma representação mecanizada da semântica e das restrições que ajudam a detectar anomalias;
- Geração automática de documentos: adaptação de *sites* que se modificam dinamicamente de acordo com as preferências dos usuários poderia ser muito útil, mas requer máquinas que acessam a representação da semântica destas fontes de informações.

A tecnologia da Web Semântica providencia novas possibilidades para a gestão do conhecimento como: pesquisas inteligentes ao invés de pesquisas utilizando palavras-chave. Neste contexto entram os metadados.

A Web Semântica utiliza metadados para dar significado aos seus recursos, estes metadados são criados através de alguns padrões próprios para web. Na próxima seção são apresentados conceito, características e padrões de metadados.

2 METADADOS

Pode-se dizer que metadados são "dados sobre dados". Neste contexto, metadados referem-se a estrutura descritiva da informação sobre outro dado, o qual é usado para ajudar na identificação, descrição, localização e gerenciamento de recursos da web. Entretanto, eles podem ser aplicados em qualquer meio.

De acordo com Iannella e Waugh (1997), no contexto da web, três aspectos devem ser considerados no desenvolvimento de metadados:

- Descrição de recursos: informação expressa através de metadados, determinado pelo objetivo e tipo do recurso.

[4] Tecnologia usada para criar as URLs.

- Produção de metadados: sumário da descrição dos dados, que pode tornar-se um processo extremamente caro e cansativo quando realizado manualmente. A tendência é realizar automaticamente esse processo, incentivados pelo uso das tecnologias XML e RDF.
- Uso de metadados: Envolve o uso e acesso de metadados, é especialmente relevante para a localização de recursos na web. Neste contexto, metadados devem incluir informações sobre os recursos, tais como a identificação, descrição, estrutura.

Outras questões de relevância relacionadas a metadados podem ser apontadas, ainda conforme Iannella e Waugh (1997):

- Devido ao grande número de padrões de metadados existentes, é possível descrever um recurso a partir de mais de um conjunto de descritores;
- Deve existir um conjunto de padrões de metadados específicos de forma a acomodar os diferentes tipos de recursos.
- A internacionalização de padrões deve ser considerada. A maioria dos padrões é baseada em descritores em inglês;
- Metadados devem ser gerados no momento em que o recurso é criado e disponibilizado na web, evoluindo à medida que o recurso é modificado. No entanto, alguns tipos de metadados específicos, tais como críticas sobre um filme ou um artigo, podem ser gerados à parte.
- Metadados são dados. Assim, apresentam também características de armazenamento e acesso, bem como dificuldades de interpretação de seu conteúdo.

2.1 Padrões de Metadados

A criação de um único padrão de metadados que aborde todas as áreas do conhecimento humano é cada vez mais difícil devido a alguns fatores, tais como: o tamanho da web, um padrão único seria composto por um número elevado de elementos descritores referentes aos diversos domínios do conhecimento; mediante a variedade de descritores, a catalogação se tornaria exaustiva e exigiria um conhecimento mais específico.

Segundo Souza, et al. (1997), os padrões de metadados têm como função fornecer as definições e formar uma rede para automatizar registros de propriedades e dados cadastrais de forma padronizada e consistente.

Existem padrões diferentes de metadados para finalidades distintas de informações. Para se ter uma idéia da variedade de esforços são apresentados alguns destes padrões: DIF

(Directory Interchance Format) – padrão para criar entradas de diretórios que descrevem um grupo de dados: GILS (Government Information Locator Service) usado para descrever informações governamentais; FGDC (Federal Data Geographic Committee) usado na descrição de dados geoespaciais; MARC (Machine Readable Cataloging) usado para a catalogação bibliográfica; CIMI (Consortium for the Interchange of Museum Information) que descreve informações sobre museus.

Dentro da categoria para descrição de recursos na web, encontra-se o Dublin Core, que apresenta uma estrutura a partir de um conjunto de descritores simples e genéricos que objetiva a descoberta e o gerenciamento de recursos na web. Também não exige conhecimento de especialistas no momento de descrever os recursos, devido à simplicidade de utilização, podendo ser usado por qualquer tipo de usuário, talvez por isso seja o padrão recomendado pela W3C[3] para utilização na web.

2.1.1 Dublin Core - DC

O padrão Dublin Metadata Core Element Set, possui uma lista central de quinze elementos descritores, sendo um dos padrões mais utilizados na web. Possui pares (nome atributo/valor atributo) como estrutura e sua utilização não possui maior complexidade.

Segundo Lagoze et al. (1996) o Dublin Core pretende ser simples e, para facilitar o uso pelos criadores e mantenedores de documentos web, descritivo o suficiente para auxiliar na recuperação de recursos na Internet.

Isto gera um paradoxo: os descritores fornecidos pelo DC, por serem genéricos e simples, não cobrem a total necessidade de descrição dos recursos, pois dependendo da aplicação a que se refere um recurso, não se encontram descritores apropriados, ou que, no mínimo, possam ser aproveitados.

Para sanar esta ausência, foram desenvolvidos elementos extras que complementam os 15 elementos definidos pelo DC, que podem ser denominados de q*ualifiers.* Estes *qualifiers* são avaliados pelo DCMI (*Dublin Core Metadata Iniciative*) para fazerem parte do conjunto de descritores às aplicações.

É dada preferência aos *qualifiers* que podem ser utilizados de maneira geral por várias aplicações. O DCMI recebe sugestões concernentes de padrões existentes adicionais que possam servir como *qualifier*s. Estas sugestões são analisadas, debatidas e aprovadas ou não pelo DCMI.

[3] URL do W3C: <http://www.w3c.org>

Para a representação destes *qualifier*s, é dada preferência aos vocabulários, anotações formais e termos mantidos e estabelecidos pelas agências já conhecidas dos usuários. Os implementadores desenvolvem *qualifier*s adicionais para uso dentro de aplicações e domínios específicos. Tais *qualifier*s podem ser reusados por outras comunidades dentro do contexto mais amplo (DCMI, 2001).

Segundo o DCMI (2001), o Dublin Core *Qualifiers* possui duas classes:

- Refinamento do Elemento: um elemento refinado compartilha o significado do elemento de uma maneira mais específica e restrita. Se um usuário não compreende o refinamento do elemento ele deve ignorar o *qualifier* e retornar ao elemento geral.
- Esquema de Codificação: identificam esquemas que auxiliam na interpretação de um elemento. Esses esquemas incluem vocabulários controlados e anotações formais ou regras para a representação do mesmo. Como esquema tem-se o Perfil de Aplicação abordado a seguir.

2.1.2 Perfil de Aplicação

Perfil de aplicação é um tipo de esquema de metadados. Este esquema consiste de elementos projetados de um ou mais esquemas de padrões de metadados, combinados por implementadores para uma aplicação particular.

Um perfil de aplicação possui as seguintes características segundo Heery e Patel (2002):

- São elaborados sob medida para aplicações particulares, ou seja, são criados elementos que melhor descrevem os dados de uma aplicação específica dentro de um domínio também específico.
- Estes elementos podem ser utilizados de um ou mais *namespace*s[4] já existentes e aplicados combinadamente em uma aplicação particular.
- O Perfil de Aplicação permite o uso dos elementos já definidos em um *namespace* existente, mas não podem criar novos elementos.

Um perfil de aplicação tem sua utilidade no que se refere à divulgação e publicação da maneira com que os pesquisadores estão utilizando os padrões de metadados. Sendo assim, desenvolve novos padrões e elementos que melhor descrevem suas necessidades de aplicação.

[4] Esquema que define unicamente todos aqueles elementos elaborados pelo registro de autoridade para um namespace particular.

Ressalta-se que é necessária a publicação e a manutenção destes padrões e elementos para a disponibilização dos mesmos a outros pesquisadores e comunidades. Esta publicação é realizada no que se denomina *Schema Registry*, que é o local (portais web, banco de dados) onde se armazena o registro de todos os esquemas associados com um *namespace* e todos os perfis de aplicação contendo os elementos associados àquele *namespace*. Para tanto, são necessárias aplicações específicas, adotando parâmetros e linguagens abordados a seguir.

3 FERRAMENTAS PARA DESENVOLVIMENTO DA WEB SEMÂNTICA

Como já foi abordado, para que a Web Semântica se torne realidade é preciso que alguns instrumentos/ferramentas sejam utilizados em conjunto, visando o entendimento homem/máquina. São elas o RDF e o XML. Segue uma breve apresentação das mesmas, ainda que o foco do artigo não esteja em aprofundar discussões acerca da aplicação e implementação destas tecnologias.

3.1 Resource Description Framework – RDF

O RDF tem por objetivo definir um mecanismo de representação de metadados para descrever recursos não vinculados a um domínio específico de aplicação. Resultado do trabalho em conjunto desenvolvido por várias comunidades.

O RDF recebeu a influência de várias fontes diferentes. As principais influências vieram das comunidades de padronização da web (HTML, XML e SGML), da Biblioteconomia (metadados de catalogação), da representação do conhecimento (ontologias), da programação orientada a objetos, da linguagem de modelagem, entre outras.

Segundo Moura (2001), na área de descoberta de recursos, o RDF possibilita a implementação de mecanismos de pesquisa mais eficientes. Na área de catalogação, o mesmo pode ser utilizado para descrever os recursos de informação em um sítio da web, como em uma biblioteca digital. Na área de agentes inteligentes o RDF pode facilitar o intercâmbio de informações e o compartilhamento de conhecimento.

O RDF é um sistema para auxílio ao desenvolvimento de metadados cuja finalidade é promover a interoperabilidade entre aplicações que compartilham informações que sejam entendidas por sistemas na web (ZANETE, 2002?). Metadados representados em RDF são usados para dar significado aos recursos da Web Semântica por permitir que estes sejam manipulados e compreendidos por máquinas.

Ele não predefine qualquer semântica nem pressupõe um domínio específico de conhecimento. Trata-se assim de um mecanismo de descrição neutro, que serve para descrever recursos de qualquer área do conhecimento (RDF, 1999 *apud* BAX; REZENDE, 2001).

A estrutura de descrição de recursos – RDF é composta por três tipos de objetos: recursos, propriedades e triplas. Um **recurso** é o que será descrito por uma expressão RDF. Todo recurso é identificado por um *URI* (*Uniform Resource Identifier*, incluindo aí o *Uniform Resource Locator* - *URL*). Uma **propriedade** é qualquer característica utilizada para descrever um recurso.

Em RDF, um domínio de conhecimento é definido via um *RDF Schema* (RDF, 1998). É no *RDF Schema*, portanto, que é definida a semântica e as características de uma propriedade. Uma aplicação que crie metadados em RDF e outra que utilize estes metadados devem utilizar o mesmo *Schema* para um funcionamento adequado.

Uma **tripla** é formada por um recurso, uma propriedade e um valor para a propriedade daquele recurso. Uma tripla possui a seguinte forma <sujeito, predicado, objeto>. O significado de uma tripla pode ser resumido como: "o recurso (sujeito) que possui a propriedade (predicado) com determinado valor (objeto)". Um valor ou objeto pode ser tanto um outro recurso quanto um tipo primitivo definido por XML (*Extensible Markup Language*) (BAX; REZENDE, 2001)

Por exemplo, a tripla <"http://www.urcamp.tche.br/josiane/metadados", "criador", "Josiane"> teria o significado: Josiane é a criadora da página <http://www.urcamp.tche.br/josiane/metadados>. É importante notar que um recurso pode ter mais de um valor para uma dada propriedade. Por exemplo, suponha que o indivíduo X e o indivíduo Y tenham construído a página http://pagina.com.br/ a existência das duas triplas < "http://pagina.com.br/","criador", "indivíduo X"> e <"http://pagina.com.br/", "criador", "indivíduo Y"> em um documento RDF não seria errônea.

Todas as triplas representam um grafo direcionado que vai do nodo sujeito para o nodo objeto e o arco tem o nome da propriedade. Um recurso é representado graficamente por uma elipse enquanto um terminal é representado por um retângulo. As triplas acima seriam representadas por:

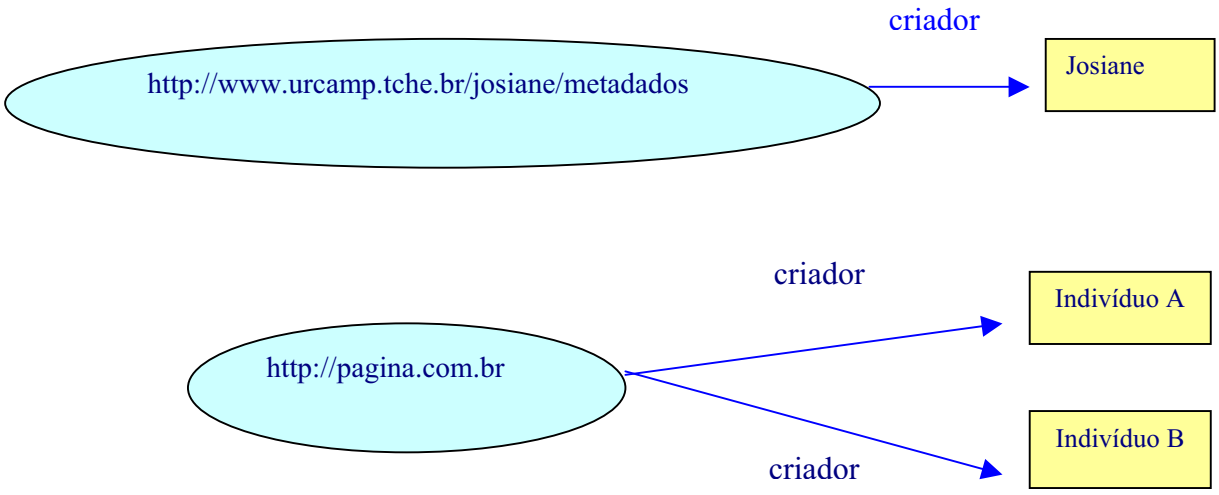

Figura 1: Declaração RDF[5]

O RDF destaca-se pela simplicidade com que busca estruturar o conteúdo contido na web. Ele não é uma linguagem, mas sim um modelo de representação para descrição semântica de recursos.

3.2 *Extensible Markup Language* - XML

O XML nada mais é do que uma linguagem de marcação de dados (metadados) que oferece aos seus usuários a descrição de dados estruturados, facilitando declarações precisas do conteúdo de documentos e mais ainda, facilitando a recuperação destes documentos via web. (DUARTE; FURTADO JUNIOR, 2000?)

De acordo com os autores supracitados, os arquivos XML são textos para leitura por uma pessoa assim como o HTML e podem "codificar o conteúdo, as semânticas e as esquematizações para uma grande variedade de aplicações desde as simples até as mais complexas, dentre elas: - um simples documento, - um registro estruturado tal como uma ordem de compra de produtos, - um objeto com métodos e dados como objetos Java ou controle ActiveX, (...) – Todos os links entre informações e pessoas na web." (DUARTE; FURTADO JUNIOR, 2000?)

A linguagem XML supre as deficiências da HTML, permitindo a criação de marcações definidas pelo próprio usuário e, desta forma, proporcionando uma maior descrição dos recursos em termos de metadados. Também fornece uma linguagem sofisticada de folha

[5] A sintaxe de RDF pode ser expressa em XML da forma especificada em RDF (1999).

de estilo – a XSL *(eXtensible Stylesheet Languag*e), baseada no padrão DSSL *(Document Style and Semantics Specification Languag*e) que adiciona estilos visuais (cores, tipos de fontes, etc.) aos documentos web.

Desta forma, a formatação do documento é tratada separadamente de sua estrutura, resolvendo assim um dos principais problemas do HTML, sem incorrer em um sistema de marcações complexas como a SGML (MOURA, 2001).

XML é a representação textual do dado. O componente básico em XML é o *element,* isto é, o texto limitado entre delimitadores (*tags*) < > ...</> (incluindo os próprios delimitadores) tal como pessoa, nome, idade e e-mail. É possível associar atributos a elementos. Um atributo é definido como um par (nome, valor).

A linguagem XML deve respeitar duas restrições: *tags* devem estar corretamente aninhadas e atributos devem ser únicos. Quando um documento atende a essas duas restrições diz-se que um documento é bem formado, sendo possível organizá-lo segundo uma estrutura de árvore e representá-lo via XML na web, o que auxilia na recuperação da informação.

4 PROJETOS SOB A PERSPECTIVA DA WEB SEMÂNTICA

Atualmente, existem projetos em desenvolvimento em várias comunidades científicas internacionais visando criar ferramentas para descrição de recursos eletrônicos, ou seja, para que os computadores consigam interpretar a linguagem humana e até mesmo indexar de forma eletrônica os documentos[6] inseridos na Internet ou nas bases de dados destas próprias comunidades científicas.

São exemplos destes projetos:

- Scorpion Project, desenvolvido pela comunidade americana;
- Projeto INDEXA, desenvolvido no Brasil;
- DESIRE - Development of a European Service for Information on Research and Education, da Comunidade Européia.

[6] Como documento, neste contexto, tem-se: artigos*, papers, surveys, páginas, e-books, links,* dentre outros suportes informacionais armazenados na Internet.

Este trabalho apresentará o Projeto Scorpion, como estudo de caso representativo dos projetos internacionais e o Projeto Indexa, representante da comunidade brasileira, no cenário de projetos para indexação eletrônica.

Estes projetos são comentados logo a seguir, a fim de ilustrar os conceitos abordados nas seções anteriores, bem como com o objetivo de demonstrar o estado da arte dos projetos de indexação automática no contexto web.

4.1 Projeto Scorpion

Desenvolvido pela *Online Computer Library Center* – OCLC[7] desde 1996. Este projeto utiliza banco de dados da *Dewey Decimal Classification* – CDD eletrônica, um dos esquemas de classificação mais utilizado pela comunidade bibliotecária, em nível internacional, para classificar materiais em bibliotecas, centros de documentação e em qualquer tipo de bases de dados informacionais.

O sistema é amplamente utilizado no mundo para classificar o conhecimento de maneira hierárquica. Foi desenvolvida por Melvil Dewey em 1786 e atualmente é um dos sistemas mais utilizados no mundo das bibliotecas (cerca de 90% das bibliotecas americanas a utilizam, ao contrário do Brasil, onde são poucas as bibliotecas que a utilizam em comparação a outros países).

O projeto Scorpion tem como objetivo a exploração da indexação e a catalogação de recursos eletrônicos. Como supracitado, tem seu foco primário na estrutura de ferramentas para reconhecimento automático do assunto em busca avançada.

Este projeto tem como vantagem em relação aos métodos tradicionais de indexação, poupar tempo e agilizar o trabalho de indexação dos bibliotecários no tratamento da informação, além de tentar fazer o que uma biblioteca faz em um espaço circunscrito (o ambiente físico).

Faz isso através de listas de cabeçalhos de assuntos controlados; buscando sinergia entre o mundo desorganizado da web e o mundo organizado das bibliotecas, através da inserção de novas tecnologias via Internet.

Como proposta principal, o Scorpion se propõe a tentar combinar a indexação e a catalogação focalizando-se na construção de ferramentas para reconhecimento de assunto automaticamente, combinando técnicas de Biblioteconomia com técnicas de recuperação automática da informação, reduzindo custos na catalogação e facilitando este processo.

[7] URL da OCLC: www.oclc.org

Como sugestão dos desenvolvedores, eles pensam na emissão de uma lista de assuntos em potencial, após o tratamento do documento pela tecnologia Scorpion, a ser apresentada a um "catalogador humano" que poderia escolher os assuntos mais apropriados desta lista retirada do texto, através de algoritmos de *hankiamento* e palavras baseadas na CDD, como descritores do documento em processo de indexação.

Esta metodologia agiliza o trabalho do indexador como também evita que o mesmo faça uma leitura técnica exaustiva do material para indexá-lo, uma vez que o Scorpion pode emitir uma lista de opções com os termos mais significativos no documento.

O projeto visa otimizar esta etapa, oferecendo ao indexador estas palavras, além de favorecer também o processo de transformação destas palavras escolhidas para representarem o documento em linguagem artificial, que para a Biblioteconomia, segundo Lancaster (1993) é a etapa de tradução em que, depois de escolhidas as palavras do texto, faz-se um controle das permitidas a serem utilizadas na base de dados, evitando duplicidade de semântica entre as escolhidas para esta representação.

Este projeto decidiu utilizar a CDD eletrônica como banco de consultas para geração de indexação automática porque a Dewey possui alto grau de integridade de classes (o que equivale dizer que para cada assunto há uma única representação, pelo menos em nível de contexto), além de ser considerada pelos desenvolvedores do projeto, como uma base de conhecimento excelente para uma ferramenta de tarefa de indexação automática[8].

A CDD oferece meios uniformes para denotação do tópico primário de um recurso assim podendo agrupar artigos semelhantes (bases de dados de artigos semelhantes para procura de artigos de comum interesse). O agrupamento na CDD é feito pela divisão em classes, sendo dividido em suas 10 grandes áreas (princípio da decimalidade), conforme apresenta-se a seguir:

000 – Generalidades
100 – Filosofia e Psicologia
200 – Religião
300 – Ciências Sociais
400 – Lingüística
500 – Ciências Exatas e Puras
600 – Ciências aplicadas, Tecnologia

[8] Convém lembrar que esta não é uma visão unânime. Maniez (1993) aponta que as classificações hierárquicas como a CDD, CDU e LC possuem claras deficiências estruturais, rupturas lógicas, alto grau de dispersão, terminologia pobre e acentuadas manifestações enumerativas.

700 – Artes e Recreação
800 – Literatura
900 – História, Geografia e Biografia

Quadro 1**:** Divisão do Conhecimento CDD das classes principais

De acordo com esta classificação hierárquica, em que cada conceito é denotado por um número que identifica sua posição na hierarquia (de no mínimo três números, conforme agrupamento acima), tem-se na CDD o armazenamento do conhecimento humano dividido por áreas principais, com subdivisões por conceitos, partindo do genérico para o específico. Quanto mais longo o número de classificação[9], maior será a especificidade representada (conceito mais específico).

O Scorpion cria o seu banco de dados relacionado com a Dewey, utilizando os registros da CDD do ESS – *Editorial Support System,* da OCLC, para criar bancos de dados de recuperação enfileirados para atribuir assuntos automaticamente aos documentos.

A atribuição dos cabeçalhos é obtida pelo Scorpion através do SMART – *System for Manipulating And Retrieving Text* baseado em representações estatísticas através dos algoritmos ATN e ATC, que computam peso aos termos, através de grau 0.0 a 1.0, para reduzir a vantagem de recuperação que registros longos têm sobre os menores.

Os registros do ESS baseiam-se na natureza hierárquica da CDD e são representadas as classes mais amplas de um número CDD no banco de dados da ESS através de registros separados. Por exemplo, na classe 000, ele abarcaria:

000 *Generalities*

005 *Computer programming, programs, data*

005.3 *Programs*

Esta relação entre classes, ou hierarquia estrutural em Dewey, mostra que todos os tópicos fazem parte dos tópicos mais amplos acima hierarquicamente e caso a busca se dê por qualquer um deles, recuperará informações semelhantes - o que expande significativamente a indexação (aumentando a revocação) -, porém sem onerar a base de dados, como é o caso de outras tecnologias de indexação eletrônica, que criam índices invertidos com todas palavras não pertencentes ao antidicionário, criado com palavras pertencentes, em sua maioria, de artigos e preposições.

[9] Mais longo, no sentido de possuir mais algarismos. Por exemplo, de acordo com a CDD, a classificação 658 é mais genérica do que a 658.3. A 658.34 é mais específica que a 658 e mais específica também que a 658.3, assim sucessivamente.

O Scorpion vai além e evita a duplicidade semântica, por seguir a CDD que possui integridade em sua estrutura, evitando conceitos idênticos representados por palavras (e notações) distintas.

O acréscimo que o Scorpion traz ao estado da arte em indexação eletrônica é justamente o uso deste acúmulo de conhecimento já consolidado pela área de Biblioteconomia. Ele, além de utilizar indexação por *hanking* estatístico de palavras, com algoritmos que computam peso aos termos, vai além, buscando interdisciplinaridade com os instrumentos para atribuição de cabeçalhos de assunto, bem como utilizando a CDD para coletar, estruturar e padronizar o tratamento destes assuntos.

O Scorpion não pode substituir a catalogação humana. Há ainda muitos aspectos que esta oferece e que são difíceis, senão impossíveis, de se automatizar. Porém, ele deve produzir ferramentas que ajudem a reduzir o custo e o tempo despendido pela indexação tradicional (humana) automatizando a tarefa de atribuir assunto quando os documentos estiverem disponíveis eletronicamente.

4.2 Projeto Indexa

Desenvolvido pela Escola de Ciência da Informação da Universidade Federal de Minas Gerais, o Indexa é um sistema de indexação de *sites* em mecanismos de busca na Internet.

Seu objetivo é criar uma ferramenta automatizada que analise o documento para o auxilio na preparação de *websites*, antes que este seja submetido à classificação e indexação pelos mecanismos de busca. Isto com a geração das descrições das páginas analisadas em RDF (como formato de representação de metadados), através de um processo semi-automático e supervisionado.

O projeto quando concluído deverá ser capaz, segundo seus idealizadores, de:

- Analisar as informações de uma página e propor modificações e ajustes (alguns automáticos e outros manuais), para auxiliar os processos de indexação usados pelos 5 maiores mecanismos de busca na web;
- Identificar fatores regionais (brasileiros) que influem no processo de classificação pelos mecanismos de busca;
- Tornar uma informação melhor recuperável.

A plenitude do projeto se dará quando os principais mecanismos de busca utilizarem padrões genéricos de descrição de recursos como o RDF. Em sua versão piloto, o Indexa utiliza para sua aplicação na web, o Java (Servlets e JPS) e o formato de representação de metadados adotado é o RDF.

As ferramentas utilizadas visam preparar *sites* para serem indexados por máquinas de busca, além de submeter URLs a estas máquinas, bem como analisar estes *sites* indexados pelos mecanismos de busca.

Para preparar *sites* a serem indexados pelos mecanismos de busca, o Indexa se propõe a selecionar palavras candidatas a bons descritores através da:

- Freqüência da palavra na página (TF – *term frequency);*
- Especificidade da palavra na coleção (inverso da freqüência da palavra na coleção da máquina de busca – IDF – Inverted Document Frequency);
- Posição da palavra na estrutura da página.

Utilizando os parâmetros acima citados seria possível analisar um documento HTML e dele extrair possíveis palavras-chave. Este procedimento calcula, para cada palavra do documento, o seu grau de relevância para a descrição da página.

Para a realização do cálculo da Especificidade é necessário obter o número de ocorrências de cada palavra no índice do mecanismo de busca e também qual o peso atribuído a cada faixa de valores de freqüência. Já o cálculo da localidade de uma palavra é realizado com a atribuição de valores a cada posição possível de ocorrência dentro de um documento. Os valores podem variar de acordo com a importância do local do ponto de vista de Recuperação da Informação.

Para submeter as URLs aos mecanismos de busca, o Indexa analisa os fatores extrínsecos que fogem do controle do web designer e como solução propõe o envio da página para vários mecanismos de busca para tentar que a quantidade de sua submissão origine vários *links*. Isso com o objetivo de que páginas com número mínimo de conteúdo tornem-se visíveis.

Para análise dos *sites* indexados, o projeto Indexa propõe a realização de relatórios automáticos periodicamente, para avaliar o desempenho de uma página nos *hankings* dos mecanismos de busca na Internet. Isto através de técnicas de mineração de dados para cálculo de *hanking* que serão adotados para a geração de regras lógicas que descrevem o comportamento dos algoritmos de relevância.

De um modo geral, o Indexa trabalha de uma forma mais efetiva do que os sistemas de busca, trabalhando os documentos (páginas) a serem submetidos aos mecanismos de busca existentes, a fim de que estes se tornem utilizados pelos usuários da Internet.

5 CONSIDERAÇÕES FINAIS

O artigo teve como epicentro discussões acerca da Web Semântica, seu conceito, objetivo e sua importância diante do cenário da necessidade de criação de métodos satisfatórios no tratamento de informações distribuídas pela Internet. Onde o advento da tecnologia Web Semântica fornecerá maior precisão ao acesso à informação relevante através da estruturação e da representação semântica dos dados.

Procurou-se dar destaque à necessidade de objetivar a formação de um ambiente coeso tanto em nível tecnológico quanto em nível de representação do conteúdo dos documentos.

Apresentou-se projetos que adotam a metodologia de implementação de acordo com a Web Semântica, bem como identificou ferramentas utilizadas para tal implementação: DC, RDF e XML. Acredita-se que estas ferramentas devam ser trabalhadas em conjunto no desenvolvimento de bases de dados e plataformas semânticas que visem coadunar homem/máquina na busca pela satisfação na organização do conhecimento das necessidades informacionais em ambiente web.

O destaque para o uso do Dublin Core vem do fato de ser este um padrão de metadados bastante simples quanto aos seus campos de preenchimento, o que facilita a padronização na utilização de recursos e tratamento da informação pelos próprios criadores de páginas web (catalogadores leigos), possibilitando uma pré-classificação destas páginas, o que facilitaria a indexação das mesmas pelas bases de dados.

Quanto aos projetos, tecendo uma breve comparação entre os existentes na Comunidade Européia e Americana, com os projetos em nível nacional (quanto ao uso da tecnologia da Web Semântica e indexação automática), percebe-se que no primeiro grupo, os

projetos parecem seguir a tendência pela busca da qualidade dos recursos de representação e recuperação de documentos de boa qualidade (especificidade).

No Brasil, parece que os projetos ainda estão bastante tímidos e muitos ainda buscam posições de destaques dos documentos indexados nos mecanismos de busca, sem a devida preocupação com a qualidade da informação, preocupando-se com o *hankiamento* dos mesmos, (apesar do projeto Indexa ser de grande importância uma vez que o mesmo oferece um canal de pré-tratamento da informação inserida na web).

Para otimizar este cenário, acredita-se ser necessário a participação de profissionais bibliotecários a fim de que seja possível aproveitar o conhecimento legado de indexação e classificação, (vide projeto Scorpion que está tentando utilizar a CDD como forma de classificar documentos de forma eletrônica, aproveitando sua integridade de classes semânticas) não somente reproduzindo tais técnicas, mas, em parceria com profissionais da tecnologia da informação, superar estas técnicas e desenvolver novas metodologias para tratamento, recuperação e organização da informação eletrônica.

A Biblioteconomia sempre tratou a informação e a organizou através de metadados descritivos como, por exemplo, uma entrada de título em um catálogo de uma biblioteca (quer este catálogo fosse eletrônico ou manual), o que faz com que o usuário não precise estar com a obra em mãos para ter idéia do que trata a mesma.

Porém, este tratamento sempre foi dado a acervos limitados fisicamente, através de uma rede uniformizada. Mas o que fazer para tratar volumes informacionais gigantescos e com crescimento exponencial como as grandes bases de dados virtuais web, uma vez que o custo pela catalogação especializada é bastante caro, além de ser uma prática dispendiosa e morosa cujas técnicas são direcionadas para o tratamento apenas de volumes informacionais relativamente circunscritos?

A web pode oferecer ferramentas automatizadas para busca da informação sem tratamento e a Biblioteconomia pode oferecer sua experiência e teoria na organização da informação. Estas comunidades, aliando suas técnicas e tecnologias, podem atrelar recursos poderosos para enfrentar problemas de acesso e manutenção à informação eletrônica de qualidade.

Esse estado revela a necessidade de melhorias no tratamento da informação no cenário brasileiro de indexação eletrônica. Inúmeras implementações foram detectadas na bibliografia de projetos disponíveis via rede, com relação à indexação por *hankiamento*, através de algoritmos de balanceamento de pesos e de estatística. Porém precisa ser trabalhada a visão semântica e ontológica se o objetivo for o de otimizar e agregar qualidade à indexação

eletrônica, almejando a relevância dos resultados nas buscas via Internet, evitando a recuperação do lixo informacional.

REFERÊNCIAS

BAX, M. P.; REZENDE, B.V. **Projeto Indexa***:* ferramenta de auxílio à divulgação de informações na web. [2001?]. Disponível em: <http://www.paradigma.com.br/artigos/artigos_04.pdf> . Acesso em: 20 outubro de 2002.

BERNERS. T.B.; HENDLER; J., LASSILA, O. The Semantic Web. **Scientific American**, maio 2001. Disponível em: <http://www.scientificamerican.com/2001/050lissueberners-lee.html> Acesso em: 08 setembro de 2002.

CRISTIANINI, G. M. S.; MORAES, J. de S. **Novas tecnologias, antigas classificações**. Disponível em: <http://intermega.globo.com/biblio_fespsp/texto_69.pdf>. Acesso em: 25 novembro de 2002.

DUARTE, O.C.M.B.; FURTADO JUNIOR, M. B. **XML***:* Extensible Markup Language. Tutorial disciplina Redes de computadores. (2000?) Disponível em: <http://www.gta.ufrj.br/grad/00_1/miguel/link1.htm> Acesso em 22 maio de 2002.

DCMI - DUBLIN CORE METADATA INITIATIVE. **Dublin Core Qualifiers**. 2001. Disponível em: <http://www.dublincore.org/documents/2000/07/11/dcmes-qualifiers>. Acesso em: 12 setembro de 2002.

FARIA, C. G.; GIRARDI, R. **Uma análise da web semântica e suas implicações no acesso à informação**. 2002. Disponível em: <http://maae.deinf.ufma.br/ensino/ia/artigos> Acesso em: 08 janeiro de 2003.

HEERY, R; PATEL, M. Application profiles*:* mixing and matching metadata schemas. 2002. **Ariadne,** n. 25. Disponível em: <http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.> Acesso em: 12 agosto de 2002.

HENDLER, J. Agents and the semantic web. **IEEE Intelligents Systems**, mar./abr. 2001.

IANNELLA, R.;WAUGH, A. **Metadata***:* Enabling the Internet. 1997. Disponível em: <ftp://www.dstc.edu.au/RDU/reports/CAUSE97>. Acesso em: 28 novembro de 2001.

IKEMATU, R. S. Gestão de metadados: sua evolução na tecnologia da informação. **Bate Byte**, Setembro 2000. Disponível em: <http:www.pr.gov.br/celepar/celepar/batebyte/edicoes/2000/bb101/gestao.htm>. Acesso em: 18 janeiro de 2002.

LAGOZE, C. et al. **The Warwick Framework** – A container Architecture for Aggregating Stes os Metadata. 1996. Disponível em: <http://dlib.org/dlib/july96/lagoze/07lagoze.html>. Acesso em: 15 setembro de 2002.

LANCASTER, F. W. **Indexação e resumos***:* teoria e prática. Brasília: Briquet de Lemos, 1993.

MANIEZ, J. L'évolution des langages documentaires. **Documentaliste-Sciences de l'information**, 1993, v. 30, n. 4-5, p. 254-259.

MILLER, E. W3C Semantic Web Activity. **Semantic Web Activity Statement**. Disponível em: <http://www.w3.org/2001/sw/Activity> Acesso em: 24 agosto de 2002.

MOURA, A. M. C. **A Web Semântica***:* fundamentos e tecnologias. 2001, Instituto Militar de Engenharia. IME. Rio de Janeiro. Disponível em: http://ipanema.ime.eb.br/ãnamoura/public/ websemantica.zip. Acesso em 12 janeiro de 2002.

PROJECT Scorpion. Disponível em: <http://orc.rsch.oclc.org:6109/> Acesso em: 12 junho 2002.

RDF: Resource Description Framework. **Model and Syntax Specification**. W3C Recommendation 22 febreary 1999. Disponível em: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222> Acesso em: 25 maio 2002.

RDF: Resource Description Framework. **Schema Specification**. 1998. Disponível em: <http://www.w3.org/TR/1998/WD-rdf-schema/ > Acesso em: 25 janeiro 2002.

SOUZA, T. B.; et al. Metadados*:* catalogando dados na Internet*.* **Transinformação**, v. 9, n.2, 1997, maio/ago. Disponível em: <http://puccamp.br/~biblio/tbsouza92.html>. Acesso em: 19 agosto de 2000.

ZANETE, N. H. **Introdução ao RDF**: Resource Description Framework. [2002?]. Disponível em: <http://www.faccar.com.br/zanete/zaneteRDF.hmt> Acesso em: 16 setembro de 2002.

WIERENGA, K., DESIRE: Development of a European Service for Information on Research and Education. **Ariadne***,* n.5, 1996, set. Disponível em: <http://www.ariadne.ac.uk/ issue5/desire/> Acesso em: 12 junho 2002.

WORLD WIDE WEB CONSORTIUM (W3C). **Resource Description Framework (RDF)** Model and syntax specification, 1999. Disponível em: <http://www.w3c.org/TR/1999/REC-rdf-syntax/> Acesso em: 16 junho de 2002.

ABSTRACT

This paper approaches the Semantic Web: a new version of web development, through projects as Scorpion and Desire. The aim of these projects in to organize knowledge stored in their files and web pages promissing the understanding of human language by the machines to recover information, without the user needs to dominate refined searching strategies. The article presents the metadatas pattern Dublin Core as the present day most used pattern by the project developer communities in the area of the Web Semantic and approaches RDF as suitable structure for the visionary of this new web to develop semantic outlines in the representation of the information made available through net, as well as XML as language of demarcation of structured data. Reveals the need of improvements in the treatment of the information in the Brazilian scenery of electronic indexation so that the same can accompany the new paradigm of recovery of information and organization of the knowledge.

KEYWORDS: Semantic web. Organization of the Knowledge. Metadatas. Dublin Core. RDF. XML. Project Scorpion. Project Indexa.

Originais recebidos em 21/06/2004