JOSÉ EDUARDO SANTAREM SEGUNDO

Recursos tecno-metodológicos para descrição e recuperação de informações na Web

Marília
2004

JOSÉ EDUARDO SANTAREM SEGUNDO

Recursos tecno-metodológicos para descrição e recuperação de informações na Web

Dissertação apresentada no Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Câmpus de Marília, para obtenção do título de Mestre.

Linha de pesquisa: Informação e Tecnologia

Orientadora: Dra. Silvana Aparecida Borsetti Gregorio Vidotti

Marília

2004

Santarem Segundo, José Eduardo
S233r Recursos tecno-metodológicos para descrição e recuperação de informações na Web / José Eduardo Santarem Segundo. – Marília, 2004.
157 f. ; 30 cm.

Dissertação (Mestrado em Ciência da Informação). – Faculdade de Filosofia e Ciências , Universidade Estadual Paulista, 2004.
Bibliografia: f. 140-150
Orientadora: Vidotti, Silvana Aparecida Borsetti Gregório

1. Internet. 2. XML. 3. Web semântica. 4. Linguagens de marcação. 5. Ferramentas de busca. 6. Descrição e recuperação de informações na Web. I. Autor. II. Título.

CDD – 004.6

JOSÉ EDUARDO SANTAREM SEGUNDO

Recursos tecno-metodológicos para descrição e recuperação de informações na Web

BANCA EXAMINADORA:

Presidente e Orientador:
Dra. Silvana Aparecida Borsetti Gregorio Vidotti
Profa. do Departamento de Ciência da Informação
Faculdade de Filosofia e Ciências
Universidade Estadual Paulista – UNESP, Campus de Marília

Membro Titular:
Dr. Marcos Luiz Mucheroni
Departamento de Ciência da Computação
Faculdade de Informática de Marília
Centro Universitário Eurípedes de Marília

Membro Titular:
Dra. Plácida Leopoldina Ventura Amorim da Costa Santos
Departamento de Ciência da Informação
Faculdade de Filosofia e Ciências
Universidade Estadual Paulista - UNESP, Campus de Marília

Local: Universidade Estadual Paulista - UNESP
Faculdade de Filosofia e Ciências
UNESP – Câmpus de Marília
Sala: Auditório da Biblioteca

Horário: 09 horas

Data: 05/02/2004

A MINHA LUCIANA, COMPANHEIRA, INCENTIVADORA, BIBLIOTECÁRIA, FUTURA MÃE DOS MEUS FILHOS E PRINCIPALMENTE MULHER...QUE FOI FUNDAMENTAL ME DANDO APOIO, AMOR E CONFIANÇA PARA QUE PUDÉSSEMOS FINALIZAR MAIS ESTE PROJETO.

AGRADECIMENTOS

A meus pais José Eduardo e Vera, por seus esforços e amor. Tudo que sou hoje é graças ao vosso empenho. Obrigado pela educação, amor e confiança... sempre!

Especialmente à Dra. Silvana Vidotti, minha orientadora, exemplo de docente e a quem admiro muito pela profissional e pessoa que é. Obrigado pelo incentivo, por me fazer amadurecer, pelas críticas, pelas madrugadas, pelos ensinamentos, pela confiança e, principalmente, por nunca ter deixado de participar desta caminhada, mesmo tendo passado por uma grande dificuldade em sua vida pessoal. Que Deus continue protegendo-a. Foi um orgulho ter sido seu orientando.

A toda minha família, pelo apoio e por entender meus momentos de ausência para que pudesse concluir esta dissertação.

À Universidade Estadual Paulista, da qual me orgulho muito de ser funcionário e aluno.

Ao professor Doutor José Augusto Chaves Guimarães e Doutora Plácida Leopoldina Ventura Amorim da Costa Santos, pelos apontamentos e contribuições dadas na qualificação, que foram de grande valia para conclusão deste trabalho, assim como pelas contribuições durante a pesquisa.

Ao Doutor Marcos Luiz Mucheroni, pelos ensinamentos e por nos fazer entender as relações do tempo com o espaço e com o cérebro.

Ao pessoal do STI, Zeca, Marco, Alexandre e estagiários que colaboraram no trabalho durante meus afastamentos para œoncluir esta pesquisa.

Ao Elvis, pelas discussões sobre XML.

À Vânia, da biblioteca, pelo incentivo.

À Sylvia do escritório de pesquisa por manter sempre o alto astral.

Ao pessoal da Pós-Graduação, principalmente à Cecília e à Maria Inês, sempre providentes.

Aos docentes da pós-graduação em Ciência da Informação, pelos ensinamentos durante as disciplinas cursadas.

Ao Departamento de Ciência da Informação e a todos seus docentes, pelas contribuições durante o trabalho.

Aos alunos da minha turma com os quais compartilhei momentos muito agradáveis durante as aulas.

Ao meu computador, companheiro inseparável das madrugadas e que nunca deixou de funcionar durante este tempo.

Ao Sr. Steffano, pela correção ortográfica.

A Deus, por todas as coisas e especialmente por ter colocado essas pessoas em minha vida.

A todos que direta ou indiretamente contribuiram na elaboração deste trabalho de pesquisa.

“Embora ninguém possa voltar atrás e fazer um novo começo, qualquer um pode começar agora e fazer um novo fim”

SANTAREM SEGUNDO, José Eduardo. *Recursos tecno-metodológicos para descrição e recuperação de informações na Web*. 2004. 157 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2004.

RESUMO

A tecnologia trouxe à Ciência da Informação uma nova partícula em seu objeto de estudo - a informação na Web; trouxe, também, uma aproximação muito grande entre as Cências da Informação e da Computação. A Internet vem crescendo rapidamente, incrementando a explosão de informações, de forma a termos uma grande quantidade de informação disponível na Web. Desse modo, torna-se necessário investigar tecnologias para descrição e recuperação de informações que possibilitem a organização da informação digital no âmbito da World Wide Web. Valendo-se de pesquisa documental em fontes das áreas de Ciência da Computação e Ciência da Informação e da própria rede Internet foram analisadas as principais linguagens e os recursos para publicação de informações na Web, as formas de descrição e recuperação de informação, as propostas de novos padrões e de estrutura de dados e abordadas as novas ferramentas que vêm sendo discutidas e implementadas, objetivando a organização da informação digital. Verificou-se o delineamento de uma Web Semântica, que se trata de uma extensão da Web atual e que propõe uma nova arquitetura, de maneira que possamos dar significado a toda informação encontrada neste novo conceito de Internet. Tais aspectos permitem concluir que a criação da Web Semântica é uma questão de tempo e que, em breve, essa nova extensão da Web passará a ser um pedaço consistente e qualificado de informações dentro da Internet, possibilitando a várias comunidades a construção de conhecimento a partir de dados confiáveis encontrados na rede.

Palavras-chave: Internet; XML; Web Semântica; Linguagens de Marcação; Ferramentas de Busca; Descrição e recuperação de informação na Web.

SANTAREM SEGUNDO, José Eduardo. *Recursos tecno-metodológicos para descrição e recuperação de informações na Web*. 2004. 157 f. Dissertation (Master Degree in Information Science) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2004.

ABSTRACT

The technology brought to Information Science a new particle in its object of study - the information in the Web; it brought, also, a very great approach enters sciences of the Information and the Computation. The Internet comes growing of frightful form, developing the explosion of Information, of form the terms a countless amount of available information in the Web. In this way, one becomes necessary to investigate technologies for description and recovery of information that make possible the organization of the digital information in the scope of the World Wide Web. Using itself documentary research in sources of the areas of Computer Science and Information Science and proper net Internet had been analyzed the main languages and the resources for publication of information in the Web, the forms of description and recovery of information, the proposals of new standards and structure of boarded data and the new tools that they come being argued and implemented, objectifying the organization of the digital information.The delineation of a Semantic Web was verified, that if deals with an extension of the current Web and that it considers a new architecture, thus let us can give meant to all information found in this new concept of Internet. Such aspects allow to conclude that the creation of the Semantic Web is a time question and that, in briefing, this new extension of the Web will inside start to be a consistent and qualified piece of information of the Internet, making possible to some communities the construction of knowledge from found trustworthy data in the net.

Keywords: Internet; XML; Semantic Web; Markup Languages; Tools of Search; Description and recovery of information in the Web.

LISTA DE FIGURAS

Figura 1 – Pesquisa no Google 95
Figura 2 – Pesquisa no Yahoo! 96
Figura 3 – Camadas da Web Semântica 111
Figura 4 – Arquitetura mais simples de camadas da Web Semântica 113
Figura 5 – Diagrama 1 - RDF 115
Figura 6 – Diagrama 2 - RDF 116
Figura 7 – Esquema RDF 119
Figura 8 – Classes RDF 120
Figura 9 – Hieraquia de Classes RDF 121
Figura 10 – Regras PICS 131

LISTA DE EXEMPLOS

Exemplo 1 - Tags 41
Exemplo 2 – Tag Title 42
Exemplo 3 – Tag Meta – Expires 43
Exemplo 4 – Tag Meta - Title 44
Exemplo 5 – Tag Meta - Description 44
Exemplo 6 – Tag Meta - Keywords 44
Exemplo 7 – Tag Meta – Content Type 45
Exemplo 8 – Tag Meta - Author 45
Exemplo 9 – Tag Meta – Identifier-URL 45
Exemplo 10 – Tag Meta – Reply To 45
Exemplo 11 – Tag Meta – revisit-after 45
Exemplo 12 – Tag Meta - Publisher 46
Exemplo 13 – Tag Meta - Generator 46
Exemplo 14 – Fechamento do cabeçalho 46
Exemplo 15 – Tag Meta - Refresh 46
Exemplo 16– Tag Meta – Window-target 46
Exemplo 17– Tag Meta - Robots 47
Exemplo 18 – Css Externo 50
Exemplo 19 – CSS Incorporado 50
Exemplo 20 – CSS Inline 50
Exemplo 21 – CSS Importado 50
Exemplo 22 – Elementos sobrepostos 55
Exemplo 23 – Elementos não sobrepostos 55
Exemplo 24 – XML Simples 62
Exemplo 25 - CDATA 64
Exemplo 26 - Elemento 64
Exemplo 27 - PCDATA 65
Exemplo 28 - ATTLIST 66
Exemplo 29 – Declarações de entidades 66
Exemplo 30 - Notation 68
Exemplo 31 – DTD Externa 69
Exemplo 32 – Link Simples 76
Exemplo 33 – Links Estendidos 77
Exemplo 34 – Declaração RDF 114
Exemplo 35 – Sintaxe serialização RDF 117
Exemplo 36 – Descrição das classes RDF 124

SUMÁRIO

Agradecimentos 4
Resumo 7
Abstract 8
Lista de Figuras 9
Lista de Exemplos 10
Sumário 11
1 Introdução 13
2 A informação na Web como objeto de estudo da Ciência da Informação 24
3 A descrição da informação na web: linguagens e recursos 32
3.1 SGML (Standard Generalized Markup Language) 35
3.2 HTML (HyperText Markup Language) 38
3.2.1 Estrutura da HTML 41
3.3 CSS (Cascading Style Sheet) 48
3.4 XML (eXtensible Markup Language) 51
3.4.1 Definição da linguagem XML 54
3.4.2 Características da Linguagem XML 56
3.4.3 Documentos com DTDs 58
3.4.4 Padrões da estrutura do XML 59
3.4.5 Documentos XML 61
3.4.6 Modelagem de Documento XML 70
3.4.7. Apresentação visual de um documento XML utilizando-se de XSL 72
3.4.8. Links poderosos em XML com Xlinks e Xpointers 73
4 A recuperação da informação na web: das ferramentas de busca às potencialidades do delineamento de uma "web semântica" 80
4.1 Formas de localização, indexação e descrição como suporte à recuperação da informação 80
4.2 Recursos para recuperação de informações 87
4.3 Formas de apresentação das informações 93
4.4 Métricas para avaliação do grau de satistação do usuário 97
4.5 Web Semântica: padrões para organização da informação digital 103
4.6 Delineamento da Web Semântica, uma nova proposta para a Web 107
4.7 Proposta de estrutura da Web Semântica 110
4.7.1 Camada de estrutura e RDF (Resource Description Framework) 113
4.7.2 Camadas Esquema 124
4.7.3 Camada Lógica 126
4.8 Novas tecnologias aplicadas aos processos de localização, descrição e recuperação da informação da Web 127
4.8.1 Web Services 127
5 Considerações Finais: um olhar para o futuro 134
Referências 140
Apêndice 152

CAPÍTULO 1
INTRODUÇÃO

1 INTRODUÇÃO

O número de usuários conectados à rede Internet vem aumentando a cada dia em todo o mundo. No Brasil, o crescimento também é exponencial, em 2000 Takahashi (2000, p.139) já afirmava que

> [...] a Anatel tinha interessante proposta no sentido de se implantar um esquema de acesso próprio à Internet [...] é a possibilidade de se apoiar decisivamente a distribuição de provedores por uma ampla região, e não somente em uma cidade, a preços fixos e uniformes; independente da distância entre o usuário prospectivo e o provedor.

Segundo IBGE (2003, p.1),

> O número de moradias com microcomputador cresceu 15,1%, de 2001 para 2002, e o de domicílios com computador ligado à Internet teve crescimento ainda mais acentuado (23,5%). Entre os bens duráveis pesquisados, a taxa de crescimento das moradias com computador foi a que mais cresceu. Em 2001, primeiro ano em que se pesquisou a existência de microcomputador nas residências, 12,6% tinham esse bem. No ano seguinte, a proporção de domicílios com computador já era de 14,2% e a daqueles ligados à internet, 10,3%.

Essa universalização da informação e o advento das novas tecnologias vêm trazendo benefícios para a sociedade, cultura e para a economia, e teve seu início algum tempo atrás, através do rádio, do telefone e da televisão.

A comunicação pela Internet por meio do uso da hipermídia, ou seja, de textos e figuras publicadas e acessadas através de links, vem derrubando barreiras de distâncias, classes sociais e idiomas.

Com o rápido crescimento da rede e da disseminação da informação, está sendo possível facilitar transações econômicas, fluxo de capitais, ofertas de bens e serviços, além da troca de informações e conhecimentos entre as pessoas.

O produto “informação” tido como base principal dos meios de comunicação e que se torna imprescindível na vida das pessoas, passa a ser fundamental na busca do conhecimento.

Santos (2002, p.106) afirma que

> [a] sociedade contemporânea se encontra em processo acelerado de transformações e rupturas, sob o signo do acesso e da detenção da informação. Tais transformações exigem definições e redefinições de métodos que valorizem e destaquem os aspectos importantes desta relação de disseminação e compartilhamento de conhecimentos e informações.

É importante ressaltar que a informação vem transformando o mundo e fazendo parte dele, como enuncia Euclides (2000, p.3)

> [...] qualquer que seja a conceituação, o que fica claro é que hoje a informação é vista como um insumo básico para o desenvolvimento de qualquer atividade. No mundo globalizado, onde o mercado necessita ser cada vez mais competitivo, constitui-se em fator determinante para o sucesso ou fracasso de uma empresa.

Definir informação é, portanto uma tarefa às vezes árdua, mas que Barreto (1998, p.168) define como

> [...] conjuntos significantes com a competência e a intenção de gerar conhecimento no indivíduo, em seu grupo e na sociedade; conhecimento esse que tem como objetivo promover o desenvolvimento do indivíduo e da sociedade em que ele vive.

É imprescindível, portanto, a disponibilização de informações de todos os tipos: de negócios, tecnológicas, cadastrais, econômicas e financeiras, políticas, legais, comerciais etc. Informações que levam à construção de conhecimento e tomada de decisões, conforme Sardenberg em Takahashi (2000, p.V) afirma:

> O conhecimento tornou-se, hoje, mais do que no passado, um dos principais fatores de superação de desigualdades, de agregação de valor, criação de emprego qualificado e de propagação do bem-estar. A nova situação tem reflexos no sistema econômico e político. A soberania e a autonomia dos países passam mundialmente por uma nova leitura, e sua manutenção – que é essencial – depende nitidamente do conhecimento, da educação e do desenvolvimento científico e tecnológico.

A Internet é hoje o meio de comunicação que permite maior interatividade entre as pessoas na troca de informações, e o seu crescimento não permite calcular estatisticamente o avanço de informação disponível. A abundância de dados digitais exacerba a mais fundamental restrição no trato da informação: os limites da compreensão humana.

> A Rede Internet pode ser entendida e visualizada como um labirinto documental no qual as informações armazenadas e apresentadas na World Wide Web (WWW ou Web) são estruturadas em sites/home-pages em forma de redes hipertextuais. As informações textuais, sonoras e imagéticas de um site possuem interligações internas e externas a outros sites determinadas pela lógica de cada fornecedor de informações ou desenvolvedor do site. (VIDOTTI, 2001,p.1)

As novas tecnologias da informação nos trazem a possibilidade de acesso a todo tipo de informação, em qualquer lugar e a qualquer momento.

No livro Cibercultura, Pierre Levy (1999, p.31) relata que

> [...] nos anos 50, Albert Einstein declarou que três grandes bombas haviam explodido durante o século 20: a bomba demográfica, a bomba atômica e a bomba das telecomunicações. Aquilo que Einstein chamou de bomba das telecomunicações foi chamado, por meu amigo Roy Ascott (um dos pioneiros e principais teóricos da arte em rede), de 'segundo dilúvio', o das informações. As telecomunicações geram esse novo dilúvio por conta da natureza exponencial, explosiva e caótica de seu crescimento. A quantidade bruta de dados disponíveis se multiplica e se acelera. A densidade dos links entre as informações aumenta vertiginosamente nos bancos de dados, nos hipertextos e nas redes. Os contatos transversais entre os indivíduos prolifera de forma anárquica. É o

> transbordamento caótico das informações, a inundação de dados, as águas tumultuosas e os turbilhões da comunicação, a cacofonia e o ceticismo ensurdecedor das mídias, a guerra das imagens, as propagandas e as contra-propagandas, a confusão dos espíritos.

Essa explosão de informações acaba gerando ansiedade entre as pessoas, e segundo Wurman (1995, p.220),

> Por muito tempo, as pessoas não se davam conta do quanto não sabiam não tinham idéia do que não sabiam. Mas hoje as pessoas sabem o quanto não sabem, e isso as deixa ansiosas.

A abundância de informação cria a pobreza de atenção e a necessidade de dirigir esta atenção de modo eficiente.

Diante de tanta informação em forma de textos, fotos, animações, áudio e vídeo existentes na Web (World Wide Web), a recuperação e organização dessas informações pelo usuário acaba dificultando a construção do conhecimento de forma estruturada.

Barité (2001, p.44) define conhecimento como

> O processo intelectual ou emocional que realiza um indivíduo para entender um fenômeno do mundo exterior e compreender seu resultado, reafirmando ou removendo sua concepção de mundo. (tradução nossa).

Para o autor "[...] o conhecimento se realiza a partir da informação, e ao socializar-se, transforma-se em informação" (p.42)

Guimarães (2000, p.9) enfoca que

> Em tempos de informação com valor estratégico, cabe criar instrumentos que se adequem a uma concepção de disponibilização de conhecimento registrado para geração de novo conhecimento, em que a vertente temática assume papel preponderante, visto resgatar a essência do conteúdo informacional.

Recuperar de forma qualitativa uma informação na Internet, tem sido hoje objeto de discussão no mundo todo. Congressos, conferências, seminários e periódicos têm permitido uma troca de conhecimentos entre as várias áreas do conhecimento.

Diante do exposto, esta pesquisa objetivou: investigar a informação na Web como objeto de estudo da Ciência da Informação e as tecnologias para descrição e recuperação de informações na Web que possibilitam a organização da informação digital.

Para tanto, foram enfocados os seguintes objetivos específicos: verificar a informação na Web como objeto de estudo da Ciência da Informação, analisar as linguagens e recursos tecno-metodológicos para descrição de informação na Internet; investigar a recuperação da informação com foco nas propostas de novos padrões de estrutura de dados, que venha facilitar a organização dos dados na Web, além de identificar as novas ferramentas que vêm sendo discutidas e implementadas, que objetivam a organização da informação na Web.

Dessa forma, pretendendo-se construir conhecimentos teóricos sobre as formas de armazenamento, descrição e recuperação de informações na Web; realizou-se um estudo descritivo e exploratório sobre a organização da informação na Internet, por meio de revisão de literatura.

O levantamento bibliográfico foi realizado em fontes primárias e secundárias da área de Ciência da Informação e Ciência da Computação, considerando as obras primárias (livros, periódicos, anais de congresso, dissertações, teses e documentos eletrônicos da Internet) e secundária (ISI Web of Knowledge, Portal de Periódicos Científicos da Capes, Probe e EBSCO OnLine). Adotou-se, como abordagem inicial para a seleção dos documentos, os critérios de pertinência com relação aos assuntos principais desta pesquisa, aos idiomas português, inglês e espanhol e período de publicação limitado aos últimos dez anos, sendo que não houve

limitação cronológica para referências citadas nos documentos selecionados.

Após o levantamento bibliográfico e a seleção dos materiais foram realizadas leituras dos textos, que possibilitaram a criação de um referencial teórico com o qual foi possível obter subsídios para um entendimento mais detalhado do processo de descrição e recuperação de informações digitais, bem como das tecnologias e padrões utilizados e/ou em desenvolvimento que visam um novo conceito de Web.

Assim, esta pesquisa apresenta no presente capítulo, uma introdução ao tema principal da pesquisa: informação na Internet, bem como os princípios metodológicos que nortearam essa investigação científica.

Analisando a utilização de tecnologias e da aproximação da Ciência da Informação com a rede Internet, no Capítulo 2 – “A informação na Web como objeto de estudo da Ciência da Informação”, descreve-se as relações da informação na Web com esta ciência, assim como a mudança de paradigma para aceitação da informação na Web como objeto de estudos.

Considerando que um dos grandes problemas a ser solucionado está relacionado ao processo de descrição automática da informação, enfocou-se no Capítulo 3 – “A descrição da informação na Web: linguagens e recursos”; analisou-se as linguagens e recursos disponíveis para publicação de documentos na Internet. Neste capítulo são abordadas as tecnologias que deram início ao processo, como a HTML (Hyper Text Markup Language), assim como as novas tecnologias que vêm sendo apresentadas para armazenamento de informações, XML (eXtensible Markup Language) e XSL (eXtensible Style Language) por exemplo, que vem possibilitando novos métodos de publicação de informação na Web.

A linguagem XML é derivada do conceito de metadados, e segundo Grácio (2003, p.112), “A descrição através de metadados proporciona, entre outras coisas, qualidade tanto para representação de um recurso, como para o resultado de uma pesquisa”. Assim, a linguagem XML é ponto fundamental de mudança na criação de Web sites e para troca de informações entre diferentes plataformas, tornando-se hoje uma das mais importantes linguagens de desenvolvimento de conteúdo para Internet.

Segundo Almeida(2002, p.6),

> [...] diversas áreas do conhecimento discutem atualmente sobre a possibilidade de melhor aproveitar a massa de informações disponível na Internet, transformando-a em algo mais gerenciável e útil. Algumas propostas em estudo contemplam a adoção da linguagem de marcação XML em conjunto com procedimentos complementares (como, por exemplo, padrões de metadados em formato eletrônico) que permitam conferir elementos semânticos à Internet.

Assim como o tempo permite o desenvolvimento de novas tecnologias, faz também com que o conteúdo de informações continue crescendo, o que pode dificultar a busca de informações pertinentes por parte dos usuários. Para minimizar os problemas da recuperação das informações, foram desenvolvidas as ferramentas de busca, que serão descritas no Capítulo 4 – “A recuperação da informação na Web: das ferramentas de busca às potencialidades do delineamento de uma Web Semântica”.

Pesquisadores, ferramentas ou mecanismos de busca são sites de busca especializados em localizar informações na Internet. O usuário digita o termo a ser procurado, geralmente uma palavra ou frase, em uma caixa em branco disponibilizada no site e, em seguida, solicita ao sistema que a busca seja efetuada. Os pesquisadores procuram a ocorrência deste termo em seus bancos de dados e apresentam os resultados na forma de

uma lista de documentos da Internet que contém a palavra ou palavras pesquisadas.

Bueno e Vidotti (1999, p.48) apontam para a importância do

> (...) acompanhamento da evolução das Ferramentas de Busca e das formas de uso de seus operadores [que] são essenciais para uma busca estratégica de informações na Internet, pois o usuário pode através delas usufruir criteriosa e conscientemente do que de melhor a WWW oferece.

Ainda, no capítulo 4, abordou-se os critérios utilizados pelas ferramentas de busca que proporcionam uma filtragem nas informações existentes na Internet e que fornecem aos usuários referências de sites que contemplam as informações que estão sendo solicitadas, através da estratégia de busca que pode conter, por exemplo: palavras chaves, partes de textos ou frases e combinações de palavras.

Apesar dessas ferramentas realizarem a busca de informações, muitas são as limitações quando se deseja realmente buscar informações de qualidade na Internet. Vários são os fatores que levam os usuários a ficarem insatisfeitos com os resultados alcançados através dos sites de busca; entre eles destaca-se a própria qualidade da informação disponível nos sites que aparecem como resultado das pesquisas nessas ferramentas.

Segundo Marcondes e Sayão (2001, p. 26),

> Um aspecto problemático da cultura de nosso tempo é o assim chamado fenômeno da explosão informacional, a grande quantidade de informações produzidas e disponibilizadas por diferentes atividades sociais, dificultando sua identificação, acesso e utilização. Na emergência da sociedade da informação, o valor desta como insumo para qualquer atividade, seja ela uma decisão econômica, um processo cultural ou de ensino/aprendizagem, uma pesquisa científica ou tecnológica, está relacionado diretamente ao seu potencial de orientar de forma econômica o dispêndio de energia para a realização desta atividade. Para que possa realizar todo este potencial, a informação relevante para um dado problema deve estar disponível no tempo certo. De

> nada adianta a informação existir, se quem dela necessita não sabe da sua existência ou se ela não puder ser encontrada.

A pouca qualidade das informações, aliada à redundância de informações disponíveis, acabam causando ao usuário um resultado muitas vezes frustrante e algumas vezes, com informações imprecisas. Durante este trabalho estaremos apresentando formas métricas para avaliar o conteúdo de informção disponível na Web, e apresentamos também uma nova forma de publicação de documentos na Web, denominada Web Semântica, onde os documentos deixam de ter caracteres apenas de apresentação e passam a ter também informações de descrição de conteúdo.

A Web Semântica é uma nova proposta encabeçada por Tim Berners-Lee, seu idealizador, que se utiliza do conceito de metadados e ferramentas como as linguagens XML e RDF para estruturar os dados, permitindo que estes sejam recuperados de forma mais precisa.

Segundo Silva e Lima (2002, p.1),

> A Web Semântica envolve o arranjo das idéias e de suas associações de forma não restrita. Assim, um computador poderia representar associações entre coisas que poderiam parecer não relacionadas, mas que de fato, compartilham algum relacionamento.

No capítulo 4, apresentamos, ainda, a Web Semântica, as técnicas e tecnologias que devem ser utilizadas para escrever documentos que possam ser recuperados por conteúdo, assim como um novo ambiente que propõe uma Web mais dinâmica e que possibilita a interoperabilidade dos dados independente de produtos (softwares e hardware) ou marcas (fabricantes).

A Web Semântica, conjuntamente com as novas tecnologias que são apresentadas, permitem que sejam apontadas e discutidas

aplicações, novas ferramentas e soluções que estão sendo desenvolvidas para a Internet.

Segundo Grácio (2002, p. 13),

> Atualmente, as mudanças tecnológicas no armazenamento e na transmissão da informação, proporcionadas pela informática e pelo avanço das telecomunicações, estão alterando a relação dos profissionais da informação com a forma de tratar essa informação, bem como dos usuários com acesso a informação armazenada.
>
> Os usuários atuais necessitam de novas informações e de novos elementos em suas pesquisas, isto é, que as base de dados atuais possuam, além de textos, elementos como sons e imagens. Devido a essas mudanças tecnológicas, eles têm, algumas vezes, o acesso à informação em tempo real e interativo, mudando assim a relação de tempo e espaço. Mas, para que o acesso à informação estocada possa atender ao usuário na sua pesquisa, a informação deve ser tratada e representada, possibilitando a sua busca e recuperação.

Finalizando o trabalho, procurou-se apontar o que a Internet nos reserva no futuro, e as novas aplicações que poderão surgir e facilitar o trabalho de todo profissional que necessite de informações no âmbito da Internet: um ambiente que poderá possibilitar uma recuperação de informações com qualidade e sem perda de tempo.

CAPÍTULO 2
A INFORMAÇÃO NA WEB COMO OBJETO DE ESTUDO DA CIÊNCIA DA INFORMAÇÃO

2 A INFORMAÇÃO NA WEB COMO OBJETO DE ESTUDO DA CIÊNCIA DA INFORMAÇÃO

A Informação é um objeto de estudo indispensável que deve ser socializada pelas unidades de informação, a fim de capacitar o cidadão ao conhecimento e à crítica, enriquecendo seu potencial, de forma a sempre se incluir na sociedade de modo sério e responsável.

Para Barreto (1994, p.3),

> A informação sintoniza o mundo. Como onda ou partícula, participa na evolução e da revolução do homem em direção a sua história. Como elemento organizador, a informação referencia o homem ao seu destino; mesmo antes de seu nascimento, através de sua identidade genética, e durante sua existência pela sua competência em elaborar a informação para estabelecer a sua odisséia individual no espaço e no tempo.

Segundo Lê Coadic (1996, p.5), “a informação é um conhecimento inscrito (gravado), sob a forma escrita (impressa ou numérica), oral ou audiovisual”.

Smith (2002, p.21) diz que

> [...] a informação pode ser definida como estruturas simbolicamente significantes, codificadas de forma socialmente decodificável e registradas (para garantir a permanência no tempo e portabilidade no espaço) e que apresentam a competência de gerar conhecimento para o indivíduo e para seu meio. Estas estruturas significantes são estocadas em função de um uso futuro, causando a institucionalização da informação.

O estudo sobre a informação e, principalmente, a importância deste estudo acabou gerando uma nova Ciência, chamada de Ciência da Informação.

Grácio (2002, p.10) aponta que

> [...] a preocupação com o tratamento da informação teve uma ênfase maior com o desenvolvimento científico e

> tecnológico ocorrido no período seguinte à Primeira Guerra Mundial, por volta de 1930, permeando o crescimento do capitalismo industrial da década de 30 e gerando um crescimento na utilização de informações de ciência e tecnologia (explosão da informação), base para o crescimento econômico da época.

Nesse contexto histórico, surgem no período de 1960 e 1970 os primeiros conceitos e definições de uma nova área, a Ciência da Informação, pautada na interdisciplinaridade, nos estudos de como tratar a informação e de como seria a atuação dos profissionais da área.

Segundo Pinheiro (1999, p.156),

> O estudo da Ciência da Informação parte do reconhecimento de sua interdisciplinaridade, de sua natureza social, forte e profundamente relacionada à tecnologia da informação e do novo papel da informação na sociedade e cultura contemporâneas, características essenciais da área.

Neste mesmo documento, Pinheiro, apresenta uma das primeiras definições, que é de Taylor (1966), posteriormente sintetizada e reelaborada por Borko (1968, p.3), em artigo em torno do que seria Ciência da Informação: “disciplina que investiga as propriedades e comportamento da informação, as forças que regem o fluxo de informação, a fim de alcançar acessibilidade e utilização ótimas”. A nova área foi por ele compreendida como um corpo de conhecimentos relacionados “à origem, coleção, organização, armazenagem, recuperação, interpretação, transmissão, transformação e utilização da informação”

Le Coadic (1996, p.56) define Ciência da Informação como “o estudo das propriedades gerais da informação (natureza, gênese e efeito), dos processo e sistemas de construção, comunicação e uso dessa informação”.

Para se definir uma ciência é importante delimitá-la, facilitando assim focar seus objetos de estudo.

Fernandes (1995, p.25) diz que

> Delimitar um objeto de estudo (como se olha) e um campo de fenômenos (para onde se olha) são parâmetros básicos para definir uma ciência e para a continuidade da atividade cientifica. A bem de alguns equívocos que explicitam tanto o objeto como os fenômenos da Ciência da Informação sem levar em conta o que seja realizar tal delimitação, a Ciência da Informação tem construído como seu objeto o que denominamos 'gestão institucional dos saberes'. Realizada por três instituições modernas: o Estado, a Ciência e o Sistema Produtivo Capitalista, esta gestão, que resulta na produção de um artefato cultural, a informação, tem por finalidade religar aquilo que por algum motivo está separado na era moderna.

Smith (2002, p.17) vai mais além dizendo que, neste início de século, devido à interação da Ciência da Informação com uma tecnologia intensa, a Ciência da Informação redefine o conteúdo e a prioridade de seus objetivos continuamente. Enunciando que

> [...] o objeto de estudo da Ciência da Informação como campo que se ocupa e se preocupa com os princípios e práticas da criação, organização e distribuição da informação, bem como com o estudo dos fluxos da informação, desde sua criação até a sua utilização, e sua transmissão ao receptor em uma variedade de formas, por meio de uma variedade de canais.

Neste contexto de definições podemos também tentar delimitar alguns pontos em Ciência da Informação, como Brasil (1998, p.72):

> O chamado setor de informação atua com dados, informação propriamente dita, conhecimento e inteligência, assim hierarquizados segundo o nível de valor sucessivamente agregado no processamento de análise, na avaliação e na contextualização para decisão. Destarte, inteligência (informações analisadas e contextualizadas para fins decisórios) e conhecimento (informações sistematizadas e assimiladas pelo indivíduo, de forma relacionada como seu saber e sentir anteriores) representam áreas multidisciplinares de alto valor agregado. Por outro lado,

> dado é um fragmento bruto e desarticulado da realidade, enquanto informação é todo conteúdo (dados, fatos, textos, sons, imagens) organizado para comunicação em qualquer formato e por meio de qualquer canal ou suporte. As questões relacionadas com dados e informações têm sido objeto freqüente dos documentos e atividades de todos os países nos seus respectivos programas para sociedade da informação, envolvendo as áreas de tecnologia da informação e comunicação, indústria e serviços de informação, como aplicação nas mais variadas áreas.

Santos (2002, p.103) diz que:

> [...] se pensarmos que hoje o contorno da economia é definido pela quantidade de informação possuída, veiculada e disseminada podemos identificar a informação como matéria-prima do mundo contemporâneo, juntamente com as tecnologias disponíveis.

A autora destaca, ainda, que

> [...] a presença da tecnologia no cotidiano das pessoas formando opinião, criando necessidades e determinando comportamentos torna a atuação do profissional de Biblioteconomia extremamente importante no processo de formação reflexiva dos sujeitos no que se refere ao uso de informações alocadas nos mais diversos suportes.

Segundo Marcondes e Sayão (2002, p.42),

> Quando se fala em informação para ciência e tecnologia, este papel é mais acentuado ainda. Isto porque a ciência institucionalizada está assentada em mecanismos de comunicação rápida dos resultados de pesquisa, que por sua vez estão hoje baseados fortemente nas tecnologias de informação.

A tecnologia e principalmente a disponibilização de informações na Web vêm contribuindo para uma mudança de paradigma e, assim, permitindo que a Internet faça parte direta e seja objeto de estudo da Ciência da Informação.

Desde a invenção do computador na década de 50, no século XX, as tecnologias de informação passaram a ser usadas pelas bibliotecas para prover acesso não só aos documentos dos seus próprios acervos, mas também aos armazenados em acervos de outras bibliotecas. (MARCONDES, 2001).

Segundo Bueno e Vidotti (2000, p. 6),

> [...] o acesso à informação eletrônica é o ponto alto das tecnologias de informática aplicadas às Bibliotecas, pois, com a tecnologia das redes eletrônicas, torna-se possível o surgimento de novos documentos e produtos e, por conseqüência, a criação de novos serviços, como a orientação aos usuários na utilização de seus recursos, o desenvolvimento de home-pages, o agendamento e o atendimento de novos serviços on-line, como a comutação, o empréstimo entre bibliotecas, a disseminação da informação e o catálogo.

Le Coadic (1996, p.109) afirma, sobre Ciência da Informação, que

> Técnicas audaciosas e os imperativos de sua tecnologia a impulsionam irresistivelmente e a fazem passar do universo do papel para o universo eletrônico. Nesse universo, informações de toda a natureza podem ser armazenadas e transmitidas sob forma digital. Após tê-las convertido, representamos qualquer texto, som ou imagem na forma de bits e bytes. Um vez digitalizadas, essas informações podem ser veiculadas por diferentes meios, nas redes de transmissão, por difusão hertziana, em (micro, mini, super) computadores e até mesmo em livros eletrônicos.
>
> Do mesmo modo, por mais diferentes que sejam os estádios de maturidade técnica e econômica dos diferentes componentes veiculados (voz, texto, imagem), o centro de gravidade das práticas informacionais desloca-se inexoravelmente de um pólo constituído pelo papel para um pólo eletrônico onde o oral e o visual retomam um lugar que o textual lhes havia tomado, deixando entrever o surgimento de uma nova cultura informacional. As gerações futuras serão mais exigentes em relação a esses meios, particularmente os audiovisuais menos formais do que os meios textuais; terão provavelmente menos tempo e estarão menos interessadas em obter informação mediante uma

> leitura constante, podendo o texto ser apresentado oralmente.

Ou seja, a relação de proximidade da tecnologia de informática com a Ciência da Informação é realmente de muita afinidade.

O computador, diferente da sua origem, quando sua principal tarefa era exclusivamente realizar cálculos matemáticos complexos, é hoje uma ferramenta importante no armazenamento, na organização, na recuperação e no intercâmbio de informações armazenadas em meio eletrônico. (BAX, 2001)

Assim, não só a tecnologia de equipamentos mas também a da utilização fundamental da rede de computadores Internet passam a ser imprescindíveis na Ciência da Informação.

Grácio (2002, p.15) afirma que

> A Internet é atualmente a expressão maior da utilização dos computadores e dos meios eletrônicos para armazenamento, a busca e a recuperação de informações armazenadas em meio eletrônico. Para recuperar a informação armazenada na Internet e transformá-la em conhecimento, são utilizadas atualmente ferramentas de busca, que consistem em programas de computadores com bancos de dados que armazenam descritores de recursos disponíveis na Internet.

Percebemos assim que a tecnologia de informática, em especial a rede Internet, está inserindo-se rapidamente na área da Ciência da Informação e que a informação disponível na Web, como suporte eletrônico de informação, cada dia está participando mais do cotidiano desta ciência interdisciplinar que utiliza a tecnologia com muita familiaridade.

> Sobre as relações interdisciplinares da Computação com a Ciência da Informação não há ‘uma completa união’ mas uma ‘co-habitação’, a coexistência das duas disciplinas, reconhecidas pelo uso do plural Ciências da Informação, com a intenção de abrigar disciplinas numa espécie de ‘guarda-

> chuva' curricular. Eles sugerem adotar Ciência da Computação e Ciência da Informação e reconhecem 'boas razões' para essa junção, pois os primeiros têm interesse em informação e tendem a 'ficar confinados ao seu papel nos sistemas de computação e envolver signos, simbolos e assim por diante (a abordagem semiótica) e seus processadores (a abordagem da informática). (PINHEIRO apud MACHLUP, MANSFIELD, 1983, p.167).

Neste capítulo, procuramos verificar como a Ciência da Informação se apresenta como área, como a informação disponível na Web passa a ser também seu objeto de estudo, mesmo como mais uma maneira de se disponibilizar a informação.

No próximo capítulo veremos as maneiras de descrever as informações na Web, através de suas linguagens e recursos.

CAPÍTULO 3
A DESCRIÇÃO DA INFORMAÇÃO NA WEB: LINGUAGENS E RECURSOS

3 A DESCRIÇÃO DA INFORMAÇÃO NA WEB: LINGUAGENS E RECURSOS.

A World Wide Web (Web ou WWW) é o serviço de informação da Internet que consiste em milhares de páginas, formadas por gráficos e fotografias, combinadas com texto. A visualização destas páginas faz-se por meio de um browser, ou seja, um programa de navegação que abre uma janela na tela do computador onde são exibidas as páginas.

A Web baseia-se no conceito de hipertexto. Nos documentos de textos tradicionais a informação é lida de forma seqüencial, ao passo que nos documentos da Web a disposição da informação é feita de uma forma fragmentada, de tal modo que se pode saltar de informação para informação, dado que as hiperligações inseridas no texto o proporcionam.

Le Coadic (1996, p.61) afirma que

> O que diferencia o conceito de hipertexto das outras formas de armazenamento eletrônico da informação é sua estrutura associativa que reproduz, muito de perto, a estrutura da memória humana e pode tornar-se seu complemento íntimo e ampliado. Permite substituir as estrutuuturas clássicas arborescentes da informação por estruturas mais ricas e mais complexas, organizadas em redes, mostrando um número infinito de caminhos, abertos a todas as navegações e interligando múltiplos objetos.

Constituída por páginas que podem ser livremente criadas e por protocolos que regulam a transferência, a Web é, com certeza, o maior repositório de informações conhecido até hoje, e é através deste mecanismo de armazenamento e recuperação de dados que comunidades, em diferentes partes do mundo, trocam informações.

Com a Web passamos a ter transferência de voz, textos, imagens, vídeos, entre outros tipos de hipermídia, além da possibilidade da informação continuar armazenada para uma consulta posterior e em diferentes equipamentos.

Para que fosse possível a existência da Web, o desenvolvimento da linguagem de marcação, em especial da HTML (HyperText Markup Language), foi fundamental.

Vale destacar que as linguagens de marcação apareceram muito antes da Internet. Uma marcação é algo que define a característica, mas não acrescenta conteúdo ao texto. Os principais formatos são: marcação de procedimento e marcação descritiva.

Hoje, a maioria dos sistemas de publicação eletrônica utilizam-se de marcação de procedimentos, preservando a apresentação do documento. Neste caso, o que é visto na tela é uma imagem real do documento, processo denominado WYSIWYG *(What You See Is What You Get,* ou seja, o que você vê é o que você obtém de resultado); as marcações ficam implícitas e normalmente são inseridas no documento através do software.

Tomemos como exemplo o Microsoft Word: ele traz em seu conteúdo textos em negrito, itálico, sublinhado, entre outras formatações; para que isso aconteça é necessário que internamente, onde o usuário não tem acesso, o texto tenha marcações para que possa realmente adquirir os formatos especificados: o usuário, assim, não consegue visualizar as marcações, vendo somente o texto já formatado.

A Microsoft utiliza-se de um sistema fechado (proprietário) de marcação e, por esse motivo os arquivos desenvolvidos no Word, não podem ser abertos em outro programa ou, quando são, apresentam falhas de configuração.

Nos sistemas de marcação descritiva não existe a preocupação com a aparência, sendo que o que realmente importa é o conteúdo. Neste caso, as marcações são explícitas e chamadas de *tags*, e se apresentam junto com o conteúdo, servindo para informar o que representa cada parte do texto: como título, capítulo, texto e figura.

Na marcação descritiva não existe a preocupação da formatação do texto quanto à apresentação do mesmo, visto que o ponto principal é definir as entidades que fazem parte do texto. Assim, em um sistema descritivo, a mídia onde o texto será divulgado não afeta o resultado final, já que não há uma preocupação com constantes de formatação e sim com o conteúdo.

A utilização de documentos eletrônicos em grande escala criou a necessidade das linguagens de marcação para facilitar a padronização dos documentos eletrônicos, o que levou algumas empresas, por volta de 1960, a definirem suas próprias linguagens de marcação, gerando uma proliferação de diversos formatos de marcação para textos.

Segundo Almeida (2002, p.6),

> Historicamente, usa-se a palavra “marcação” para descrever anotações ou marcas em um texto, que tem por objetivo dar instruções ao desenhista ou datilógrafo sobre a maneira como uma parte do texto deveria ser representada. Como exemplos, pode-se citar um sublinhado ondulado que indicaria negrito, símbolos especiais para passagens a serem omitidas ou impressas com uma fonte especial, dentre outras. Como a formatação e a impressão de textos se tornaram automatizadas, o termo foi estendido para todos os tipos de códigos de marcação em textos eletrônicos. Todos os textos impressos são codificados com sinais de pontuação, uso de letras maiúsculas e minúsculas, regras para a disposição do texto na página, espaço entre as palavras, etc. Estes elementos são um tipo de “marcação”, cujo objetivo é ajudar o leitor na determinação de onde uma palavra termina e onde outra começa, ou identificar características estruturais (por exemplo, cabeçalhos) ou simples unidades sintáticas (por exemplo, parágrafos e sentenças). Codificar ou “marcar” um texto para processamento por computadores é também um processo de tornar explícito o que é conjetural. Indica como o conteúdo do texto deve ser interpretado.

Diante de inúmeros formatos disponíveis foi criado no início dos anos 1980 um comitê chamado de “Computer Languages for the Processing of Text” que, em 1986, publicou o padrão internacional ISO

8879 (International Organization for Standardization), definindo uma linguagem padrão de marcação para textos denominada Standard Generalized Markup Language (SGML).

3.1 SGML (STANDARD GENERALIZED MARKUP LANGUAGE)

A linguagem SGML descreve um padrão para o uso de marcações descritivas mescladas ao documento. Não se trata de um conjunto pré-determinado de marcas, mas sim de uma linguagem para se definir quaisquer conjuntos de marcas, uma linguagem auto-descritiva, uma vez que cada documento SGML carrega consigo sua própria especificação.

A SGML também fornece um método padrão para nomear as estruturas de um texto, definindo modelos hierárquicos para cada tipo de documento produzido. A linguagem SGML obriga os elementos descritos, como capítulo, título e parágrafo, a se ajustarem na estrutura lógica e previsível de seu documento.

Há diferentes estruturas para cada tipo de documento: boletins informativos, manuais, catálogos, especificações de projeto, relatórios, cartas, ofícios e memorandos. Por ser um padrão internacional, a SGML permite a criação de documentos independente do sistema de computação (hardware e software) utilizado; isto significa que se pode trocar informações entre usuários em diferentes sistemas e plataformas sem nenhuma alteração.

A SGML é uma linguagem padrão de descrição de texto, utilizada largamente para codificar documentos de uso geral como livros e jornais, mas que não foi desenvolvida para utilização em documentos hipertextuais.

Segundo Guimarães (2002),

> SGML foi criada no final da década de 60 pelos pesquisadores da IBM C. Goldfarb, E. Mosher e R. Lorie, com o objetivo de construir um sistema portável (i.é, independente de sistema operacional, formatos de arquivos, etc) para o intercâmbio e manipulação de documentos.

O autor afirma, ainda que

> Duas linguagens de marcação baseadas em SGML e largamente utilizadas são: DocBook (projetada para marcação de documentação técnica) e TEI - Text Encoding Initiative (projetada para marcação de textos literários).

Um documento SGML pode ser dividido em 3 camadas: estrutura, conteúdo e estilo.

A estrutura é que define como ocorre a organização da informação em um documento; isso é feito principalmente através de um arquivo chamado DTD (Document Type Definition, ou seja, definição do tipo de documento). As definições explícitas dentro de um DTD vão desde a especificação de quais caracteres podem ser utilizados como limites das anotações, até a definição dos tipos de características, como o limite máximo do tamanho dos identificadores das anotações.

A estrutura tem as seguintes características:

- normalmente, comum a todos os documentos SGML em um dado ambiente;
- pode ou não fazer parte do documento (no caso da sua ausência será usado um padrão por omissão - o padrão definido na própria norma);
- fornece detalhes precisos de como a SGML será aplicada ao documento;
- define os caracteres que serão usados para distinguir as anotações do texto (por exemplo: <, >, />);

- define o conjunto de caracteres (ASCII - American Standard Code for Information Interchange, EBCDIC - Extended Binary Coded Decimal Interchange Code, ou outro) a ser utilizado.

De acordo com Godoy (2000),

> Um DTD também especifica regras para a relação entre os diversos elementos, por exemplo: um título de capítulo deve ser o primeiro elemento ao se iniciar um novo capítulo; ou cada lista deve conter um mínimo de dois ítens. Essas regras, definidas pelo DTD, asseguram que o documento tenha uma estrutura consistente e lógica.

A camada de conteúdo é a informação propriamente dita, ou seja, a mensagem que quem cria a informação quer passar a quem vai receber o documento. O conteúdo é escrito com tags de marcação que definem o que representa cada parte do conteúdo, como por exemplo: parágrafos, listas, tabelas e gráficos.

A camada de estilo em um documento SGML é a forma de apresentação do documento. A norma de criação da linguagem não definia os estilos a serem aplicados aos documentos, porém, em 1996 foi aprovado o formato DSSSL (Document Style Semantics and Specification Language) como padrão de estilo para ser utilizado com a linguagem SGML.

O formato DSSSL apresenta recursos para padronizar o estilo da sintaxe e da semântica e o layout de documentos SGML ou para fragmentos de documentos SGML. Todas as especificações de estilo em DSSSL são feitas descrevendo o resultado da formatação final, sem o uso de algoritmos. É o DSSSL que irá nos permitir especificar como os dados serão apresentados nos dispositivos de saída, sem entretanto, criar o formato destes dados.

Entre as vantagens da utilização da linguagem SGML para criação de textos podemos destacar: longevidade da informação, aumento

da produtividade, compartilhamento de informações e portabilidade dos dados.

A linguagem SGML, após sua criação, passou a ser utilizada largamente nos mais variados setores econômicos e acadêmicos; através dela, aproveitando-se de partes de seu conceito, que Tim Berners-Lee, um pesquisador do CERN (Conseil Europeen pour la Recherche Nucleaire – hoje European Organisation for Nuclear Research), com sede em Genebra, na Suíça, criou uma linguagem de hipertexto para sua aplicação e deu início a conhecida linguagem de marcação HyperText Markup Language (HTML).

Como a SGML era uma linguagem complexa, Berners-Lee criou um modelo de documento mais simples e compacto para documentos de hipertexto. Os idealizadores, preocupados principalmente com a apresentação visual das informações, conceberam uma linguagem para interligar computadores dos laboratórios e de outras instituições de pesquisas e para exibir documentos científicos de forma simples e fácil de acessar. Como essa linguagem de marcação era bem simples e fácil, foram rapidamente criados softwares para leitura da linguagem, assim como desenvolvidos vários documentos utilizando os novos códigos.

3.2 HTML (HYPERTEXT MARKUP LANGUAGE)

A idéia principal da linguagem HTML era que ela deveria ter a finalidade de formatar um documento para apresentação, de forma simples e útil. Sendo assim, foram ignorados vários conceitos de codificação, o que permitiu que a HTML fosse de certa forma uma estrutura quase sem regras, e com flexibilidade de identificação de início e de final dos blocos de comando, pois estes nem sempre são necessários.

Essa facilidade de publicação de documentos em HTML proporcionou um aumento exponencial no número de páginas/home-pages disponíveis na Internet.

Um grande marco, que certamente determinou o crescimento e a disseminação da informação na Web, foi a criação de um programa chamado Mosaic (por Marc Andreessen), que permitiu o acesso às informações disponíveis na Web em um ambiente gráfico. Vale destacar que antes do Mosaic só era possível a exibição de textos.

O Mosaic iniciou com a versão 0.1 em março de 1993 e foi recebendo atualizações até ter sua ultima versão, a 3.0, em janeiro de 1997.

Segundo Lima (2000),

> Graças às inovações introduzidas pelo Mosaic, o usuário passou a contar com um programa de visualização que permitiu a apresentação de textos, imagens e gráficos de uma forma atraente como a de uma página de revista. ´O aparecimento do Mosaic foi uma transformação histórica para nós que já utilizávamos a Internet´, recorda Ivan de Moura Campos, secretário de Informática do Ministério da Ciência e Tecnologia. ´Antes o acesso era muito difícil, coisa para iniciados mesmo.´

E o autor destaca ainda que:

> Em conseqüência desse avanço, a Rede Mundial viveu uma verdadeira explosão. Ao fim do primeiro ano de existência do Mosaic o número de usuários da WWW havia se tornado seis vezes maior. E o mais impressionante: no segundo semestre de 93, o número de hosts (pontos ligados à Internet com ofertas de serviços) comerciais havia ultrapassado pela primeira vez o de acadêmicos. Era a demonstração de que a Internet deixava definitivamente os círculos científicos para ganhar o mundo.

O Mosaic de Marc Andreessen deu início ao desenvolvimento e aparecimento de uma grande quantidade de softwares para leitura de

informações no formato de hipertexto, chamados Browsers ou Navegadores e, juntamente com estes softwares, apareceu a idéia de navegar ou surfar na Internet.

Os navegadores rapidamente foram ganhando versões cada vez mais sofisticadas e incorporando novas tecnologias, porém, utilizando a linguagem HTML como conteúdo padrão.

Os browsers são formatadores de HTML. Quando um documento HTML é carregado em um browser, ele lê as informações HTML e formata o texto e as imagens na tela de acordo com as instruções, através de texto, contidas no documento, que podem incluir vários níveis de cabeçalho, listas, menus, formatação de estilos de texto, etc; ou seja, o browser interpreta um conjunto de instruções e apresenta o documento segundo estas instruções.

A especificação da linguagem HTML vem evoluindo desde a sua criação; o HTML+ foi apresentado em 1993 e incluiu diversas características inéditas, como tabelas, formulários para preenchimento, figuras com legendas, e as primeiras idéias para suportar fórmulas matemáticas. Em 1994 foi desenvolvida a versão 2.0 da HTML e sua especificação foi reescrita para dar consistência e usabilidade. A versão Draft 3.0 de HTML surgiu em março de 1995, enquanto que a versão atual, recomendada pelo World Wide Web Consortium (W3C), é conhecida como HTML 4.01 e foi criada para atualizar as versões 4.0 e 3.2, as quais proviam um conjunto de características novas (Java, fluxo de texto ao redor de imagens e super/subscritos), mantendo, porém, compatibilidade com a versão 2.0 anterior.

Os documentos em HTML são como arquivos ASCII comuns, que podem ser editados em qualquer editor simples, podendo ser utilizados inclusive o “Bloco de Notas” (Notepad) do Windows ou o “vi” do Unix. As extensões para os arquivos HTML são: .htm ou .html.

Hoje em dia, existem muitos editores gráficos de HTML que permitem ao usuário configurar a página da maneira que deseja graficamente, e o software se encarrega de estruturar o script HTML.

3.2.1 ESTRUTURA DA HTML

A linguagem HTML usa seqüências de caracteres denominados tags que informam como um documento deve ser apresentado. As tags descrevem a aparência e apresentação de um documento.

Vale destacar que, como existem vários navegadores ou browsers de fabricantes e versões diferentes, alguns destes softwares podem não entender uma tag específica, o que culmina na desconsideração desta tag que acaba prejudicando a apresentação do documento.

Algumas tags são utilizadas isoladamente, outras em pares. As tags utilizadas em pares identificam o final de sua ação através do nome da tag precedida pelo caractere barra (/). As tags podem ser digitadas em letras minúsculas ou maiúsculas. Os formatos sintáticos das tags são:

```
<tag [opções]>
<tag [opções]> informação </tag>
```

EXEMPLO 1 - TAGS

Um documento HTML é delimitado pelas tags <HTML> e </HTML> e dividido em cabeçalho e corpo. O cabeçalho é delimitado pelas tags <HEAD> e </HEAD> e o corpo pelas tags <BODY> e </BODY>.

A <HEAD> contém informações sobre o documento; o elemento <TITLE>, por exemplo, define um título, que é mostrado no alto da janela do browser. Podemos utilizar, por exemplo, da seguinte maneira:

```
<HEAD><TITLE>Web Semântica</TITLE></HEAD>
```

EXEMPLO 2 – TAG TITLE

Todo documento Web deve ter um título; esse título é referenciado em buscas pela rede, dando uma identidade ao documento, assim como aparece na barra de título do browser. Sugere-se que os títulos dos documentos sejam significativos, evitando-se, portanto, títulos como “Introdução".

Além do título apresentado pela tag <TITLE>, a <HEAD> contém outras informações como as tags <META> que podem ser recuperadas por robôs de pesquisa na Internet; esses campos de informação facilitam a descrição e classificação do documento em catálogos de busca, entre outras aplicações.

A tag <META> é uma das poucas do HTML que não tem função de apresentação, ou seja, para o usuário que visualiza apenas as informações da página, é indiferente sua utilização pelo programador; talvez este seja o motivo principal que conduz os desenvolvedores de conteúdo para Web a não utilizarem os recursos disponíveis por este comando.

Poucos desenvolvedores de recursos para Web sabem que as tags <META> são utilizadas pelos sistemas de busca/ferramentas de busca (principalmente os internacionais) para descrição e indexação destas informações nas suas bases de dados e que, por meio delas, os usuários fazem a recuperação das informações.

O ideal é que as tags <META> sejam inseridas em todas as páginas HTML do website, ou no mínimo na página principal do site denominada de default.asp, index.htm, index.html, etc., sempre colocadas entre os comandos <HEAD>.

A tag <META> é normalmente especificada com os campos “NAME” ou “HTTP-EQUIV”, que definem o tipo de controle para as META tags, sendo que o campo "NAME" é utilizado para todos os tipos de controle que não correspondem as chamadas HTTP. Porém, muitas vezes, alguns robôs de busca não fazem distinção entre os valores em "NAME" ou “HTTP-EQUIV”. Juntamente com os campos “NAME” ou “HTTP-EQUIV” aparece o campo “CONTENT” que define um valor para o controle especificado.

Abaixo um exemplo típico comentado:

```
<HEAD>
<TITLE>ARMAZENAMENTO, DESCRIÇÃO E RECUPERAÇÃO DE INFORMAÇÕES NA WEB: USO DE NOVAS TECNOLOGIAS </TITLE>
<META HTTP-EQUIV="Expires" CONTENT="12/12/2003">
```

EXEMPLO 3 – TAG META – EXPIRES

O controle “expires” é usado para informar ao Browser quando o documento será considerado expirado. Indiretamente também pode ser usado para controlar o diretório Cache do usuário. Exemplo: Caso o documento tenha sido expirado, o Browser fará uma nova solicitação ao site e incluirá a nova página na área de Cache.

A área de Cache é a área onde o computador armazena as páginas visitadas já visitadas e de acordo com o tempo de visita da página, o browser não faz novamente um pedido ao site indicado, buscando a informação que está armazenada dentro do próprio computador, agilizando o processo de busca e exibição da informação.

Se em “content” for utilizado um valor ilegal, por exemplo 0 (zero), isso fará com que o Browser sempre faça um novo "request" e insira a nova página. Também serve para informar aos softwares robôs dos sistemas de buscas a validade do site e automaticamente o

apagamento do registro no sistema de busca e, eventualmente, um agendamento de uma nova visita ao site. O formato da data tem de respeitar o padrão RFC 850 (Standard for interchange of USENET messages).

```
<META NAME="title" CONTENT="ARMAZENAMENTO, DESCRIÇÃO E RECUPERAÇÃO DE INFORMAÇÕES NA WEB: USO DE NOVAS TECNOLOGIAS">
```

EXEMPLO 4 – TAG META - TITLE

O controle “title” é utilizado para informar o nome do site.

```
<META NAME="Description" CONTENT="ESTE SITE É UM TESTE PARA DISSERTAÇÃO DE MESTRADO DE JOSÉ EDUARDO SANTARÉM SEGUNDO">
```

EXEMPLO 5 – TAG META - DESCRIPTION

O controle “description” é utilizado para fazer uma breve descrição sobre o conteúdo das informações do site.

```
<META NAME="Keywords" LANG="pt-br" CONTENT="ARMAZENAMENTO, DESCRIÇÃO E RECUPERAÇÃO DE INFORMAÇÕES NA WEB, WEB SEMÂNTICA, FERRAMENTAS DE BUSCA, SGML, HTML, XML">
```

EXEMPLO 6 – TAG META - KEYWORDS

O controle “keywords” é utilizado para descrever as palavras-chave sobre os assuntos que são tratados no site. As palavras-chave devem ser separadas por vírgulas e, apesar de opcional, é importante para as ferramentas de busca a informação do idioma utilizado: para tanto, utiliza-se o campo LANG, sendo que no caso acima utilizamos “pt-br” para português do Brasil, mas poderíamos, se fosse o caso, utilizar “en-us” para americamo, “pt” para português, “it” para italiano, “fr” para francês, entre outros.

```
<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=ISO-8859-1">
```

EXEMPLO 7 – TAG META – CONTENT TYPE

O controle “content-type” define o código ISO de sua página (ISO-8859-1 corresponde ao alfabeto latino com carateres acentuados). A definição do charset é regulamentada de acordo com a língua e conjunto de caracteres utilizados pelo país de origem do idioma utilizado no documento.

```
<META NAME="Author" LANG="pt-br" CONTENT="JOSÉ EDUARDO SANTARÉM SEGUNDO">
```

EXEMPLO 8 – TAG META - AUTHOR

O controle “author” descreve o nome do autor do site.

```
<META NAME="Identifier-URL" CONTENT=http://www.unesp.br/mestrado_ci>
```

EXEMPLO 9 – TAG META – IDENTIFIER-URL

O controle “identifier-URL” define o caminho completo para o endereço do seu site.

```
<META NAME="Reply-to" CONTENT="mestrado_CI@marilia.unesp.br">
```

EXEMPLO 10 – TAG META – REPLY TO

O controle “Reply-to” define o endereço de email para contato.

```
<META NAME="revisit-after" CONTENT="10">
```

EXEMPLO 11 – TAG META – REVISIT-AFTER

O controle “revisit-after” serve para indicar um número de dias que os motores de pesquisas devem utilizar para recadastrar o site.

```
<META NAME="Publisher" CONTENT="PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - UNESP">
```

EXEMPLO 12 – TAG META - PUBLISHER

O controle “Publisher” mostra o editor do site.

```
<META NAME="Generator" CONTENT="NOTEPAD">
```

EXEMPLO 13 – TAG META - GENERATOR

O controle “Generator” mostra a ferramenta utilizada para desenvolver o site.

```
</HEAD>
```

EXEMPLO 14 – FECHAMENTO DO CABEÇALHO

Além destes ainda existem outros controles que servem para ser aplicados com a tag <META> como:

```
<META HTTP-EQUIV="Refresh" CONTENT = "3 ;URL= http://www.marilia.unesp.br">
```

EXEMPLO 15 – TAG META - REFRESH

O controle “Refresh” especifica o tempo que o Browser usará para atualizar a leitura, opcionalmente pode ser usado para redirecionar para outra página/site.

```
<META HTTP-EQUIV="Window-target" CONTENT="_top">
```

EXEMPLO 16– TAG META – WINDOW-TARGET

O controle “Window-target” especifica o nome da janela em que a página atual está aberta; também pode ser usada para evitar a abertura de novas páginas em novas seções do Browser.

```
<META NAME="ROBOTS" CONTENT="INDEX">
```

EXEMPLO 17– TAG META - ROBOTS

Informa aos "robôs" de alguns sistemas de buscas, que estes devem indexar apenas a primeira página (atributo INDEX) ou que devem continuar e indexar todas as páginas (atributo FOLLOW). Outros atributos são: "NOINDEX, NOFOLLOW".

Conforme podemos verificar nos controles acima, a tag <META> tem realmente informações importantes sobre o site e que auxiliam os sistemas de busca.

Além do cabeçalho temos o corpo da página que é definida pelas tags <BODY> e </BODY>.

Tudo que estiver contido em <BODY> será mostrado na janela principal do browser, sendo apresentado ao usuário/leitor. <BODY> pode conter cabeçalhos, parágrafos, listas, tabelas, links para outros documentos e imagens.

A própria tag <BODY> tem alguns atributos de apresentação que são aplicados ao documento como cor do fundo ou imagem do fundo.

A parte chamada de corpo de uma página HTML vai conter todas as informações que serão vistas pelo usuário.

Muitas são as tags que podem ser utilizadas dentro da área chamada de corpo de uma página HTML. No apêndice A apresentamos uma tabela com as principais tags HTML.

3.3 CSS (CASCADING STYLE SHEET)

Conforme abordado anteriormente, a linguagem HTML caracteriza-se principalmente por dar formato de apresentação às informações dispostas na Web; na maioria das tags encontramos atributos de alinhamento, alterações de cor e de estilos para textos, tabelas entre outros formatos de dados. Mesmo considerando várias tags para apresentação, as limitações da linguagem HTML levaram a criação da linguagem Cascading Style Sheet (CSS), que é uma linguagem usada para definir estilos.

A CSS foi criada no final de 1996 e era utilizada somente pelo software Internet Explorer 3 da Microsoft, ainda sem a especificação da W3C. Hoje em dia a CSS encontra-se na sua segunda versão (CSS2) e é totalmente suportada pelos browsers Internet Explorer e Netscape nas versões superiores a 4.0.

Na sintaxe CSS os nomes e valores das propriedades são listados dentro de uma chave seguida do critério de seleção para tal estilo, no qual se determina em quais elementos o estilo será aplicado. Os estilos trabalham associados aos documentos HTML, mas são escritos em uma linguagem independente.

O layout de uma folha de estilo é muito fácil de ser definido e conforme Messaros (2000):

> É preciso apenas conhecer um pouco da linguagem HTML e possuir noções básicas dos termos usados em publicação eletrônica. Como exemplo, para ajustar a cor das letras de um elemento 'H1' para azul, basta fazer:
>
> H1 { color: blue }
>
> Este exemplo mostra o que é uma 'regra' simples em CSS. Uma regra é composta de duas partes principais: um selector ('H1') [um 'string' que define a quais elementos uma regra deve ser aplicada] e uma declaração ('color: blue') [uma propriedade ('color', por exemplo) e o valor atribuído a ela ('blue', por exemplo)]. Por sua vez, esta

> declaração também possui duas partes: uma propriedade ('color') e seu valor ('blue'). Embora este exemplo especifique apenas uma das várias propriedades necessárias para montar um documento HTML, ela constitui por si só uma 'folha de estilo'. Quando for combinada com outras folhas de estilo ela determinará a apresentação final do documento (uma característica fundamental é que as folhas de estilo podem ser combinadas).

As folhas de estilo tem quatro métodos de interagir com as páginas HTML e cada um desses quatro métodos tem um nome e afeta as páginas de uma maneira diferente. Os métodos são:

- externo: onde a folha de estilos CSS fica em um arquivo separado e, então, a página HTML faz um link para esse arquivo, permitindo assim que várias páginas HTML utilizem-se do arquivo CSS especificado.
- incorporado: neste método, as especificações são realizadas no cabeçalho do documento HTML e serão incorporadas somente na página onde estão.
- inline: a especificação do método neste caso acontece dentro da própria tag HTML, ou seja dentro do próprio corpo de conteúdo HTML; neste caso somente a tag especificada é que receberá a formatação.
- importado: este método é utilizado quando a folha de estilos a ser utilizada está em uma outra área na Internet; neste caso fazemos um link para a folha de estilos sem precisar copiá-la.

A definição dos estilos citados anteriormente são realizadas desta maneira:

Externo:

```
<link rel="stylesheet" href="meu_estilo.css" type="text/css">
```

EXEMPLO 18 – CSS EXTERNO

A definição acima deve estar entre as tags <HEAD> e <HEAD> do documento html.

Incorporado:

```
<STYLE Type="text/css">
P {
  Background-color: #FFFFFF;
  Font-family:'Verdana';
  Font-size: 12 pt;
 }
</STYLE>
```

EXEMPLO 19 – CSS INCORPORADO

Neste caso utilizamos a tag <STYLE> para defini-lo. A tag <STYLE> deve estar também entre as tags <HEAD> e </HEAD>.

Inline:

```
<A HREF='pagina.htm' style='color:yellow; text-weigth:bold;' >
```

EXEMPLO 20 – CSS INLINE

Neste caso, ao invés de colocar-se uma tag <STYLE>, utilizamos uma definição dentro da tag <A>, utilizando o comando style apenas como atributo.

Importado:

```
<STYLE type="text/css">
 @import url (http://www.teste.com.br/outroestilo.css);
</STYLE>
```

EXEMPLO 21 – CSS IMPORTADO

A definição acima, no método importado, também deve ser feita entre as tags <HEAD> e </HEAD>.

Dentre as vantagens de utilização da Linguagem CSS juntamente com a HTML destacamos: os diferentes estilos podem ser aplicados no mesmo documento - permitindo que o autor redirecione seu conteúdo para novos documentos; a fácil manutenção do documento – é muito mais ágil modificar uma simples página de estilo que todo o documento HTML; e a consistência do documento - a uniformidade do arranjo é um aspecto importante do desenho do Website e o CSS pode garantir que todos os documentos tenham o mesmo desenho e arranjo.

3.4 XML (eXTENSIBLE MARKUP LANGUAGE)

Com a utilização efetiva da linguagem de marcação SGML para produção de textos eletrônicos e da HTML para publicação de informação na Web, percebeu-se que a HTML não supria as necessidades reais na Internet, principalmente pelo fato de que a linguagem HTML estava preocupada principalmente com a aparência do documento. Isso conduziu ao desenvolvimento da linguagem XML que mescla a confiabilidade e segurança da linguagem SGML com a tecnologia Web para publicação de informação.

Implementações industriais na linguagem SGML mostraram a qualidade intrínseca do formato estruturado em árvore dos documentos XML.

A Linguagem XML foi construída a partir de um subconjunto da SGML e é otimizada para distribuição através da Web, sendo definida pelo W3C, assegurando que os dados estruturados serão uniformes e independentes de aplicações e fornecedores, ou seja, um documento

desenvolvido em uma plataforma computacional (ex: Windows) pode ser lido em qualquer outra plataforma computacional (ex: Linux).

A XML é também similar à HTML em vários aspectos, pois são linguagens expressas em arquivos de texto puro (ASCII), concebidas especialmente para armazenar e transmitir dados/informações. Como representantes do paradigma das linguagens de marcação, tratam de textos com marcas embutidas que qualificam cada unidade de informação (também referidas como entidades, elementos, ou objetos) contida no texto.

Conforme Ray (2001, p.11),

> Muitos documentos codificados com HTML hoje se baseiam tanto na formatação pura, que não podem ter sua finalidade redefinida com facilidade. Apesar disso, a HTML foi um passo brilhante para a WEB e um pulo gigante para as linguagens de marcação, pois fez com que o mundo se interessasse em documentação e vínculo eletrônico.

A grande diferença da linguagem XML para HTML é que a XML não propõe um número fixo de marcas. Um elemento XML pode ser marcado da forma que o autor do documento bem entender, ou seja, com o termo que melhor descreve a informação na opinião do desenvolvedor ou de uma comunidade específica.

A Linguagem XML passa, assim, a ser uma nova alavanca no conceito de informação na Internet, em especial na Web, pois pode proporcionar aos documentos estruturas sólidas e bem definidas que permitam a localização destes em diferentes locais.

Segundo Bryan (1998),

> A linguagem XML é o resultado do trabalho de um grupo de especialistas estabelecido em 1996 pelo W3C com o objetivo de propor uma simplificação de SGML que fosse voltada às necessidades específicas da Web.

De acordo com Ray (2001, p.VII), “a XML não substituirá a HTML; na verdade, a HTML está sendo absorvida gradativamente pela XML, tornando-se uma versão mais clara de si mesma”.

Segundo W3 Consortium (2003),

> A linguagem XML, embora baseada na linguagem HTML, foi projetada justamente para executar melhor a tarefa de gerenciamento de informação exigida pelo crescimento exponencial das informações na Internet. O formato de um documento XML possibilita essa atividade, pois expressa de uma maneira simples e padrão, a delimitação das informações do documento, facilitando, assim, a transmissão e o processamento dos dados nele inseridos e propondo a integração com tecnologias não proprietárias. (tradução nossa)

Para Bax (2001, p.37),

> Pode-se dizer que a passagem de uma marcação estrutural com HTML para uma marcação semântica com XML é uma fase importante no esforço para se transformar a Web de um espaço global de informação em uma rede universal de conhecimento.

A XML permite agregar semântica aos documentos, deixando por conta de cada aplicação, a interpretação da marcação atribuída a este conteúdo. Esta abordagem amplia significativamente as possibilidades do uso das linguagens de marcação, entre elas a capacidade de definir metadados – dados que descrevem dados. (CAMPOS; SANTACHE; TEIXEIRA, 1999)

Além da maneira simples de representar as informações do ambiente, a XML ainda tem um mecanismo prático de descrever os dados no documento, isto é, um documento XML, além de carregar os dados em si, aborda conjuntamente a descrição desses dados. Esta característica faz de uma aplicação XML um ótimo modo de compartilhar as informações com outras aplicações via Internet, pois os dados de uma aplicação

podem ser trocados entre usuários e outras aplicações em uma plataforma de maneira independente.

Através da natureza auto-descritiva da XML, aplicações que rodem em um ambiente distribuído podem fazer a renderização visual de dados marcados em XML sem um conhecimento prévio das informações contidas nesse documento.

Bax (2001, p.34) diz ainda que,

> Assim, esse paradigma permite tratar cada unidade de informação como um objeto (ou entidade) ao qual se pode atribuir características específicas, o que possibilita maior estruturação da informação. De um monte de caracteres estáticos, dispostos em uma página à espera de uma interpretação humana (o computador está longe de entender texto livre), a informação passa a poder ser interpretada e tratada automaticamente por computador.
>
> Os dados se transformam em objetos qualificados com atributos. Tem-se então, a possibilidade de reutilização automatizada da informação; pode-se mais facilmente compartilhá-la com outros usuários, organizá-la em bancos de dados e realizar pesquisas automáticas.

3.4.1 DEFINIÇÃO DA LINGUAGEM XML

Um documento XML tem a estrutura de uma árvore hierárquica, sendo formado através de um texto (em formato Unicode) com tags de marcação (markup tags) e outras informações.

Segundo Ramalho (2002, p.8),

> [...] com a XML, você pode usar um arquivo-texto para armazenar dados. Dada as suas características, esse arquivo pode ser usado para a troca de informações entre sistemas incompatíveis. A maioria dos bancos de dados consegue gerar arquivos XML como resultado de buscas e pesquisas. Portanto, um arquivo XML gerado por um bando de dados Oracle em um sistema operacional Unix pode ser compartilhado com um banco de dados Access rodando com sistema operacional Windows 98. Esse arquivo também pode

ser visualizado por uma página Web ou por outros aplicativos que suportam XML.

Os documentos XML são sensíveis a letras maiúsculas e minúsculas e sempre bem estruturados com tags de início e fim, sendo que as tags de elemento tem de ser apropriadamente posicionadas. Os elementos não podem se sobrepor. Um exemplo de sobreposição é o seguinte:

```
<title>Descrição das ferramentas de busca da Web<sub> Google, Yahoo! e Altavista </title> Maria da Silva</sub>
```

EXEMPLO 22 – ELEMENTOS SOBREPOSTOS

E, corrigindo o erro:

```
<title>Descrição das ferramentas de busca da Web <sub> Google, Yahoo! e Altavista</sub> <author> Maria da Silva</author> </title>
```

EXEMPLO 23 – ELEMENTOS NÃO SOBREPOSTOS

Caracteres especiais podem ser digitados usando referências de caracteres Unicode. Exemplo: “&” = &.

A linguagem XML permite que desenvolvedores definam seus próprios elementos, através da criação de DTD’s específicas, deixando por conta de cada aplicação a interpretação da marcação atribuída a este conteúdo.

As DTD’s ajudam a validar os dados quando a aplicação que os recebe não possui internamente uma descrição do dado que está recebendo. Mas as DTDs são opcionais e os dados enviados com uma DTD são conhecidos como dados XML válidos. Um analisador de documentos pode checar os dados que chegam, analisando as regras contidas na DTD para ter certeza de que o dado foi estruturado corretamente. Os dados

enviados sem DTD são conhecidos como dados bem formatados. Nesse caso, o documento pode ser usado para implicitamente se auto-descrever.

Os browsers podem fazer duas verificações principais em um documento XML: se ele está bem formatado ou formado e se ele é realmente válido.

Conforme Holzner (2001, p.9),

> Para ser bem formado, ele precisa seguir as regras de sintaxe estabelecidas para a XML pelo W3C na especificação XML 1.0 (que você poderá encontrar em http://www.w3.org/TR/REC-xml). Informalmente, bem formado significa que o documento precisa conter um ou mais elementos e um deles, o elemento raiz, precisa conter todos os outros elementos. Cada elemento também precisa ser aninhado corretamente dentro de quaisquer elementos delimitadores.

O autor afirma ainda que “Um documento XML é válido se houver uma definição de tipo de documento (DTD) associada a ele e se o documento for compilado com essa DTD.” (p.9)

3.4.2 CARACTERÍSTICAS DA LINGUAGEM XML

A linguagem XML provê uma representação estruturada dos dados que mostrou ser amplamente implementável e fácil de ser desenvolvida, além de um padrão que pode codificar o conteúdo, as semânticas e as esquematizações para uma grande variedade de aplicações, desde as mais simples até as mais complexas; dentre elas destacamos:

- Um simples documento;
- Um registro estruturado tal como uma ordem de aquisição de produtos;

- Um objeto com métodos e dados como objetos Java ou controles ActiveX;
- Um registro de dados. Um exemplo seria o resultado de uma consulta a um banco de dados;
- Apresentação gráfica, como interface de aplicações de usuário;
- Entidades e tipos de esquema padrões; e
- Todos os links entre informações e pessoas na Web.

Uma característica importante é que uma vez tendo sido recebido o dado pelo computador-cliente, tal dado pode ser manipulado, editado e visualizado sem a necessidade de relacionar o computador-servidor. Dessa forma, os servidores têm menor sobrecarga, reduzindo a necessidade de computação e reduzindo também a requisição de banda passante para as comunicações entre computadores cliente e servidor.

A XML é considerada de grande importância na Internet e em grandes intranets, porque provê a capacidade de interoperação dos computadores por ter um padrão flexível, aberto e independente de dispositivo. As aplicações podem ser construídas e atualizadas mais rapidamente e também permitem múltiplas formas de visualização dos dados estruturados.

A mais importante característica da XML se resume em separar a interface com o usuário (apresentação) dos dados estruturados. A HTML especifica como o documento deve ser apresentado na tela por um navegador, enquanto a XML define o conteúdo do documento. Por exemplo, em HTML são utilizadas tags para definir tamanho e cor de fonte, assim como a formatação de parágrafo. No caso da XML são utilizadas as tags para descrever os dados, como por exemplo: tags de assunto, título, autor, conteúdo, localização, referências e datas.

A XML ainda conta com recursos tais como folhas de estilo definidas com eXtensible Style Language (XSL) e Cascading Style Sheets (CSS), para a apresentação de dados em um navegador.

A XML separa os dados da apresentação dos dados de processo, o que permite visualizar e processar o dado de diversas formas, utilizando, para tanto, diferentes folhas de estilo e aplicações.

Essa separação dos dados da apresentação permite a integração dos dados de diversas fontes. Informações de consumidores, usuários, compras, ordens de compra, pagamentos, catálogos, acervos bibliográficos, resultados de busca, podem ser convertidos para XML no *middle-tier* (espécie de computador-servidor), permitindo que os dados sejam trocados online tão facilmente como as páginas HTML mostram dados hoje em dia. Dessa forma, os dados em XML podem ser distribuídos através da rede para os diversos computadores-clientes.

3.4.3 Documentos com DTDs

Como anteriormente mencionado, na XML as regras que definem um documento são ditadas por DTDs (Document Type Definitions), as quais ajudam a validar os dados.

Com os dados XML válidos e com os bem-formatados, o documento XML se torna auto-descritivo porque as tags dão idéia de conteúdo e estão misturadas com os dados. Devido ao formato do documento ser aberto e flexível, pode ser usado em qualquer lugar onde a troca ou transferência de informação é necessária.

Desta forma, podemos usar a XML para descrever informações sobre páginas HTML, ou descrever dados contidos em objetos ou regras de negócios, transações eletrônicas comerciais e bibliotecas digitais, por exemplo.

A XML pode ser inserida em documentos HTML, conforme definido pelo W3C como *data-islands*. Esse recurso permite que um documento HTML possa ter múltiplas formas de visualização, quando se faz uso da informação semântica contida no documento XML.

Segundo Ramalho (2002, p.31),

> [...] modelar adequadamente os dados e validar o seu conteúdo é uma preocupação que deve ter prioriadade na sua lista de tarefas. A estrutura dos dados pode ser chamada de esquema ou até mesmo de vocabulário. Ao definir um esquema, você pode criar regras para que os dados de um documento obedeçam a determinadas regras preestabelecidas.

O autor afirma, ainda, que:

> A DTD define a estrutura do documento, assim como especifica uma lista dos elementos válidos e seu possível conteúdo. Quando se trabalha sozinho, tem a liberdade de definir seus elementos do jeito que quiser. Contudo, como a XML é uma linguagem cuja principal finalidade é compartilhar dados, você deve pensar nas outras pessoas que estão envolvidas no processo de criação e manipulação dos dados. Usando a DTD voê pode criar um modelo de dados que pode ser usado por outros grupos de pessoas envolvidas no processo de trodar dados. Por meio da DTD, você pode verificar se os dados recebidos de terceiros são válidos e se seguem as normas definidas.

3.4.4 PADRÕES DA ESTRUTURA DO XML

A XML é baseada em padrões de tecnologia comprovadamente otimizados para a Web. Os padrões que compõem a XML são definidos pelo W3C (World Wide Web Consortium) e são os seguintes:

- EXtensible Markup Language (XML) - é uma Recomendação que é vista como o último estágio de aprovação do W3C. Isso significa que o padrão é estável e pode ser aplicado à Web e utilizado pelos desenvolvedores de ferramentas.

- XML Namespaces - é também uma Recomendação a qual descreve a sintaxe de namespace, ou espaço de nomes, e que serve para criar prefixos para os nomes de tags, evitando confusões que possam surgir com nomes iguais para tags que definem dados diferentes. Segundo Holzner (2001, p.61),

> Namespaces permitem garantir que um conjunto de tags não poderá entrar em conflito com outro. Namespaces permitem que você inclua um nome seguido por um sinal de dois-pontos antes dos nomes de tag e atributo, mudando esses nomes para que não entrem em conflito.

- Document Object Model (DOM) Level 1 - é uma Recomendação que provê formas de acesso aos dados estruturados utilizando scripts, permitindo aos desenvolvedores interagir e computar tais dados consistentemente. DOM é uma API (Applications Programming Interface) independente de plataforma e linguagem que é utilizada para manipular as árvores do documento XML (e HTML também).

- EXtensible Stylesheet Language (XSL) - é atualmente um rascunho e apresenta duas seções: a linguagem de transformação e a formatação de objetos. A linguagem de transformação pode ser usada para transformar documentos XML em algo agradável para ser visto, assim como transformar em documentos HTML, e pode ser usada independentemente da segunda seção (formatação de objetos). A Cascade Style Sheet (CSS) pode ser usada para XML simplesmente estruturado mas não pode apresentar informações em uma ordem diferente de como ela foi recebida.

- XML Linking Language (XLL) - e XML Pointer Language (XPointer) - são também rascunhos. A XLL é uma linguagem

de construção de links que é similar aos links da HTML, porém com mais recursos, uma vez que os links podem ser multidirecionais e podem existir em nível de objetos e não somente em nível de página. Já a Xpointer é um tipo de link parecido com os links internos do html, porém com recursos adicionais, sem precisar de marcas nos alvos.

Os esquemas modelados para XML descritos pelo W3C em XML-Data Note - e Document Content Description for XML (DCD) - estão ainda sendo desenvolvidos.

3.4.5 DOCUMENTOS XML

> A XML é inerentemente hierárquica, ou seja, um documento XML é definido como uma forma de uma árvore hierárquica simples, em que cada documento tem um nó raiz, chamado de entidade do documento ou raiz do documento. Todos os nós são denominados elementos. Assim, um documento XML é formado por um elemento raiz e por outros elementos secundários (Anderson et al., 2001, p.568).

Os elementos podem conter dados de caractere que são conteúdos de texto associados ao elemento ou a um valor de atributo. Cada elemento de um documento XML pode ser caracterizado por atributos que são informações anexadas ao elemento; estes atributos são formados por um conjunto de pares nome-valor. A seguir apresentamos um exemplo de um documento XML simples:

```
<?xml version="1.0"?>
<bibliografia>
  <livro preço=”R$ 130,00” capa=”encadernaçao”> XML Teoria e Prática
      <autor>José Antonio Ramalho</autor>
      <editora> Berkeley </editora>
```

```
        <ano>2002</ano>
    </livro>
    <livro preço="R$ 30,00" capa="brochura">Conhecimento e aprendizagem na
                                                    nova mídia
        <autor>Pedro Demo</autor>
        <editora>Editora Plano</editora>
        <ano />
     </livro>
</bibliografia>
```

EXEMPLO 24 – XML SIMPLES

Neste exemplo, alguns detalhes sobressaem:

- O documento começa com uma instrução de processamento: <?xml ...?>. Esta é a declaração XML. Embora não seja obrigatória, a sua presença explícita identifica o documento como um XML e indica a versão da XML com a qual ele foi escrito.
- Não há declaração do tipo do documento (DTD). Diferentemente da SGML, a XML não requer uma declaração de tipo de documento. Entretanto, uma declaração de tipo de documento pode ser fornecida; além disso, alguns documentos irão precisar de uma para serem entendidos sem ambigüidade.
- Elementos vazios (<ano/> neste exemplo) têm uma sintaxe modificada. Enquanto a maioria dos elementos em um documento envolve algum conteúdo, elementos vazios são simplesmente marcadores onde alguma coisa ocorre (um separador horizontal para a marca em <hr> em HTML, por exemplo). O final ***/>*** na sintaxe modificada indica a um programa que processa o documento XML que o elemento é vazio e uma marca de fim correspondente não deve ser procurada. Visto que os documentos XML não requerem uma

> declaração de tipo de documento, sem esta pista seria impossível para um analisador XML determinar quais marcas são intencionalmente vazias e quais teriam sido deixadas vazias por um erro ou por falta de informação.

O documento XML suavizou a distinção entre elementos declarados como EMPTY e elementos que meramente não têm conteúdo. Em XML é válido usar uma marca de elemento vazio para qualquer um destes casos. Também é válido usar um par de marcas início-fim para elementos vazios: <ano></ano>.

Os documentos XML são compostos de marcas e conteúdos. Existem seis tipos de marcações que podem ocorrer em um documento XML: elementos, referências a entidades, comentários, instruções de processamento, seções marcadas e declarações de tipos de documento.

Os elementos são a mais comum forma de marcação. Delimitados pelos sinais de menor e maior (<,>), a maioria dos elementos identifica a natureza do conteúdo que envolve. Alguns elementos podem ser vazios, como visto anteriormente; neste caso eles não têm conteúdo. Se um elemento não é vazio, inicia com uma marca de início - <elemento>, e termina com uma marca de término - </elemento>. Por exemplo: <autor>José Antonio Ramalho</autor>

Atributos são pares de valores nomeados que ocorrem dentro das marcas de início após o nome do elemento. Por exemplo: <livro preço=”R$ 130,00”> é um elemento livro cujo atributo preço possui o valor R$ 130,00. Em XML, todos os valores de atributos devem estar entre aspas.

Em um documento, uma seção CDATA instrui o analisador para ignorar a maioria dos caracteres de marcação. Considere um código-fonte em um documento XML. Ele pode conter caracteres que o analisador XML

iria normalmente reconhecer como marcação (< e &, por exemplo). Para prevenir isto, uma seção CDATA pode ser usada.

```
<![CDATA[*p = &q;b = (i <= 3);]]>
```

EXEMPLO 25 - CDATA

Esta marcação de CDATA permite que o parser não identifique as informações até encontrar os caracteres “]]”, entendendo, assim, que o caractere “<” dentro da fórmula não seja entendido como marcação.

Declarações de tipos de elementos identificam os nomes dos elementos e a natureza do seu conteúdo. Uma declaração de tipo de elemento pode ser assim exemplificada:

```
<!ELEMENT  livro(autor+, editora, ano?)>
```

EXEMPLO 26 - ELEMENTO

Esta declaração identifica o elemento nomeado como livro. Seu modelo de conteúdo segue o nome do elemento e define o que um elemento pode conter. Neste caso, um livro deve conter ‘autor’ e ‘editora’ e pode conter ‘ano’. As vírgulas entre os nomes dos elementos indicam que eles devem ocorrer em sucessão. O sinal de adição após autor indica que ele pode ser repetido mais de uma vez, mas deve ocorrer pelo menos uma vez. O ponto de interrogação após ano indica que ele é opcional (pode estar ausente ou ocorrer somente uma vez). Um nome sem pontuação, como editora, deve ocorrer somente uma vez, dentro de um elemento livro.

As declarações para todos os elementos usados em qualquer modelo de conteúdo devem estar presentes para que um processador XML verifique a validade do documento.

Além dos nomes de elementos, o símbolo especial #PCDATA é reservado para indicar um conjunto de caracteres. A cláusula PCDATA significa um dado analisável, que é um texto puro sem qualquer formatação.

Os elementos que contêm somente outros elementos são denominados elementos com conteúdo de elementos. Os elementos que contêm outros elementos e #PCDATA são denominados elementos com conteúdo misturado (elementos e texto puro sem marcação). Por exemplo, a definição para livro poderia ser:

```
<!ELEMENT livro (#PCDATA | autor)*>
```

EXEMPLO 27 - PCDATA

A barra vertical indica um relacionamento "ou" e o asterisco indica que o conteúdo é opcional (pode ocorrer, nenhuma, uma ou várias vezes); por esta definição, portanto, livro pode conter zero ou vários autores. Todos os modelos de conteúdo misturado devem ter esta forma: #PCDATA que deve ser o primeiro elemento; todos os elementos devem ser separados por barras verticais e o grupo inteiro (livro) deve ser opcional.

Outros dois modelos de conteúdo são possíveis: EMPTY indica que o elemento não possui conteúdo (e, conseqüentemente, não tem marca de término) e ANY indica que qualquer conteúdo é permitido. O modelo de conteúdo ANY é, algumas vezes, útil durante a conversão de documentos, mas deveria ser evitado ao máximo em um ambiente de produção, pois desabilita toda a verificação do conteúdo deste elemento.

Declarações de listas de atributos identificam quais elementos podem ter atributos, quais atributos esses elementos podem ter, quais valores os atributos podem suportar e qual valor é o padrão (por definição – default). Uma declaração de lista de atributos pode ser assim definida:

```
<!ATTLIST livro
referencia ID #REQUIRED
preço CDATA #IMPLIED
capa ( encadernação | brochura ) 'brochura'>
```

EXEMPLO 28 - ATTLIST

Neste exemplo, o elemento livro possui três atributos: referencia, que é um ID (que significa uma referência única para cada livro, como por exemplo o número de tombo) e é obrigatório; preço, que é uma cadeia de caracteres e não é obrigatório; e capa, que deve ser dura ou brochura, sendo que o padrão é brochura. Cada atributo em uma declaração tem três partes: um nome, um tipo e um valor padrão.

Declarações de entidades permitem associar um nome com algum outro fragmento de conteúdo. Essa construção pode ser uma parte de texto normal, uma parte de uma declaração de tipo de documento ou uma referência a um arquivo externo que contém texto ou dados binários. Declarações de entidades típicas são exibidas a seguir:

```
<!ENTITY UNESP-MAR "Faculdade de Filosofia e Ciências">
<!ENTITY UNESP-Regimento SYSTEM “/regimento/regimento.xml”>
<!ENTITY UNESP-Logo SYSTEM "/imagens/unesplogo.gif" NDATA GIF87A>
```

EXEMPLO 29 – DECLARAÇÕES DE ENTIDADES

Existem basicamente três tipos de entidades. De acordo com Ramalho (2002, p.45), “as entidades podem ser declaradas como: internas, externas ou parâmetros.”

As Entidades Internas associam um nome com uma cadeia de caracteres ou texto literal. A entidade <!ENTITY UNESP-MAR “Faculdade de Filosofia e Ciências”> do exemplo anterior é uma entidade interna. Usando &UNESP-MAR; na confecção do documento, o usuário irá ver "Faculdade de Filosofia e Ciências" naquele local. Entidades internas

permitem definir atalhos para textos freqüentemente digitados ou textos a serem alterados, como o estado de revisão de um documento. Entidades internas podem incluir referências para outras entidades internas, mas geralmente elas não devem ser recursivas.

A especificação XML pré-define cinco entidades internas:

- < produz o sinal de menor: <
- > produz o sinal de maior: >
- & produz o E comercial: &
- ' produz um apóstrofo: '
- " produz aspas: "

As Entidades Externas associam um nome com o conteúdo de um outro arquivo. Se este contém texto, o conteúdo do arquivo externo é inserido no ponto de referência e analisado como parte do documento referente. Dados binários não são analisados e podem somente ser referenciados em um atributo; eles são usados para referenciar figuras e outro conteúdo não-XML no documento.

A entidades <!ENTITY UNESP-Regimento SYSTEM “/regimento/regimento.xml”> e <!ENTITY UNESP-Logo SYSTEM "/imagens/unesplogo.gif" NDATA GIF87>, do exemplo, são entidades externas.

O uso de &UNESP-Regimento; inserirá o conteúdo do arquivo /regimento/regimento.xml no local da referência da entidade. O processador XML analisará o conteúdo deste arquivo como se ocorresse literalmente no local.

A entidade UNESP-Logo também é uma entidade externa, mas o seu conteúdo é binário. A entidade UNESP-Logo pode ser usada somente como o valor de um atributo ENTITY (ou ENTITIES). O processador XML

passará esta informação para a aplicação, mas ele não tentará processar o conteúdo de /imagens/unesplogo.gif.

As entidades parâmetro são identificadas pelo caracter “%” antes do nome da entidade.

Segundo Ramalho (2002, p.46),

> As referências a uma entidade parâmetro são expandidas na declaração de tipo de documento e seu conteúdo passa a fazer parte da declaração. É importante lembrar-se de que as entidades parâmetro não são reconhecidas no corpo de um documento, apenas na seção DTD.

Declarações de notação identificam tipos específicos de dados binários externos. Estas informações são passadas para a aplicação de processamento, que pode fazer o uso que desejar. Uma declaração de notação pode ser assim exemplificada:

```
<!NOTATION GIF87A SYSTEM "GIF">
```

EXEMPLO 30 - NOTATION

Se presente, a declaração de tipo de documento deve ser a primeira informação de um documento depois de comentários e instruções de processamento opcionais.

A declaração de tipo de documento identifica o elemento raiz do documento e pode conter declarações adicionais. Todos os documentos XML devem ter um elemento raiz único que contenha todo o conteúdo do documento. Declarações adicionais podem vir de uma DTD externa, chamada de subconjunto externo, ou podem ser incluídas diretamente no documento, o subconjunto interno, ou ainda em ambas versões como no exemplo a seguir:

```
<?XML version="1.0" standalone="no"?>
<!DOCTYPE capitulo SYSTEM "dtdlivro.dtd" [
<!ENTITY %ulink.module "IGNORE">
```

```
<!ELEMENT ulink (#PCDATA)*>
<!ATTLIST ulink
xml:link        CDATA  #FIXED "SIMPLE"
xml-attributes CDATA  #FIXED "HREF URL"
URL             CDATA  #REQUIRED>
]>
<capitulo>...</capitulo>
```

EXEMPLO 31 – DTD EXTERNA

Este exemplo referencia um documento DTD externo - dtdlivro.dtd e inclui declarações de elementos e atributos para o elemento ulink no subconjunto interno. Neste caso, ulink dá a semântica de um link simples da especificação XLink.

As declarações no subconjunto interno não levam em conta as declarações no subconjunto externo. O processador XML lê o subconjunto interno antes do externo e a primeira declaração tem precedência.

A fim de determinar se um documento é válido, o processador XML deve ler a declaração de tipo de documento inteira (ambos os subconjuntos). Mas para algumas aplicações, a validação pode não ser precisa e pode ser suficiente para o processador ler somente o subconjunto interno. No exemplo anterior, se a validade não é importante e a única razão para ler a declaração de tipo de documento é identifcar a semântica de ulink, então a leitura do subconjunto externo não é necessária.

Estas informações podem ser comunicadas na "declaração de documento standalone". A declaração de documento *standalone, standalone="yes"* ou *standalone="no"*, ocorre na declaração XML. Um valor "yes" indica que somente declarações internas precisam ser processadas. Um valor *"no"* indica que ambas as declarações, interna e externa, devem ser processadas.

3.4.6 MODELAGEM DE DOCUMENTO XML

Como mencionado, todo processo de desenvolvimento de aplicações depende da utilização de um vocabulário que defina as informações e as regras que conduzem estas informações dentro do ambiente sintético.

A criação de um vocabulário XML depende do conhecimento que se tem da área específica na qual se deseja modelar o documento, para que seja feita, então, a criação correta de documentos estruturados que representem o ambiente virtual adequado.

Esse processo de modelagem inclui a compreensão da estrutura e do significado das informações presentes nos documentos, a especificação dos documentos - que é a tradução da análise de requisitos do assunto de pesquisa em um conjunto de regras ou esquema para criar os documentos, e a notação dos esquemas - técnica para documentar a especificação do documento, de modo que ele se torne acessível tanto para o software de processamento quanto para outros desenvolvedores que trabalharão com o mesmo documento.

Técnicas de modelagem de dados e engenharia de software podem ser utilizadas nesse processo, pois o resultado final do projeto estará totalmente relacionado e dependente de uma modelagem correta das necessidades e características do problema inicial.

A importância da modelagem de informações no projeto de um vocabulário XML surge da necessidade de se alcançar definições absolutamente precisas sobre os dados que farão parte dos documentos processados pelo sistema e a maneira eficaz de comunicação entre os usuários; o modelo de informações define o significado dos dados.

Diferentemente de aplicações tradicionais da Internet, o novo conceito diz que os dados devem ser estruturados de maneira a

possibilitar a identificação do seu conteúdo. O documento XML vem ao encontro dessas características, uma vez que o projetista tem à sua disposição a flexibilidade do projeto e da modelagem dos dados dos documentos dentro de uma mesma tecnologia.

Depois da modelagem das informações que comporão os documentos XML, ou seja, após o projeto do vocabulário outros aspectos da aplicação necessitam ser analisados. A definição da arquitetura do sistema pode gerar um modelo cliente/servidor, no qual o cliente faz requisições ao servidor que, a partir de uma base de dados, gera um documento XML que é remetido à aplicação cliente para interpretação, ou distribuído tanto para a aplicação cliente como para a servidora, que podem gerar e/ou ler os documentos XML.

Conforme Megginson (1999),

> Estas rotinas de acesso aos documentos XML, como acessar uma parte do documento e editar suas informações, podem ser utilizadas por ferramentas de manipulação de XML que farão uso de métodos oferecidos pela API do XML que oferece um conjunto de objetos e interfaces para a manipulação de documentos XML, o Document Object Model (DOM). Além do DOM, o W3C mantém, também, uma API mais simples para XML, o SAX.

W3C DOM WORKING GROUP (2002) afirma que

> O DOM oferece uma visão de documento estruturado em árvore. Uma aplicação parser, desenvolvida para acessar um documento XML através do DOM, carrega todo o documento para a memória e tem à disposição uma visão de todos os objetos na memória como numa árvore. As principais estruturas das entidades do documento são nós na árvore objeto. Acessar esses nós e manipulá-los é uma questão de navegar na árvore usando as interfaces DOM. (tradução nossa)

Quando o documento a ser analisado é muito extenso, carregá-lo todo na memória pode comprometer a performance do sistema; assim, é melhor carregar as partes do documento conforme a necessidade de uso.

Neste caso, indica-se a utilização do Simple API for XML (SAX), pois ele fornece este tipo de acesso a documentos XML.

De qualquer maneira, aplicações tanto servidoras quanto clientes, fazem uso das interfaces do DOM ou do SAX para manipular e gerar documentos XML.

3.4.7. APRESENTAÇÃO VISUAL DE UM DOCUMENTO XML UTILIZANDO-SE DE XSL

Conforme abordado anteriormente, os documentos HTML utilizam-se de documentos CSS para personalizar a apresentação visual. No caso do XML isso não é diferente, uma vez que para apresentação dos documentos formatados com tecnologia XML devem ser utilizados os recursos de folhas de estilo de formatação visual de dados.

Para Ray (2001, p.117),

> As folhas de estilo oferecem muita flexibilidade em termos de apresentação visual de uma página, além de facilitarem a manutenção de um site por intermédio da centralização das alterações de diversas partes da página em um único lugar. Além disso, ela expande as possibilidades de formatação de uma página, pois permite, entre outras coisas, mudar um atributo que funciona para todo um documento.

Em XML um dos recursos bastante utilizado para folha de estilos é o XSL (Extensible Styles Language), que se divide em duas partes: linguagem de transformação (XSLT) e linguagem de formatação (FO).

Na linguagem de formatação (FO), a aplicação da formatação é realizada direto na linha de comando do documento XML, ou seja, as tags de formatação são aplicadas ao documento conforme a necessidade, sendo apresentadas durante a escrita do documento, misturando-se com as marcas de elementos, o que difere da linguagem de transformação

que cria um estilo separado da aplicação XML e aplica ao documento XML, gerando um novo documento formatado.

Segundo Holzner (2001, p.491),

> A linguagem de transformação permite transformar documentos em outras formas, enquanto a linguagem de formatação formata e estiliza documentos de várias maneiras. Essas duas partes da XSL podem funcionar de modo bastante independente e pode-se pensar na XSL como duas linguagens e não uma. Na prática, normalmente, trasforma-se um documento antes de formatá-lo, pois o processo de transformação permite incluir as tags que o processo de formatação exige. De fato, esse é um dos principais motivos para o W3C aceitar a XSLT como o primeito estágio no processo de formatação.

O autor diz, ainda, que

> XSLT é uma especificação relativamente nova e ainda está em desenvolvimento de diversas maneiras. Existem alguns processadores XSLT, mas lembre-se de que o suporte oferecido pelo software publicamente disponível ainda não é muito forte. Alguns pacotes oferecem suporte completo para XSLT, no entanto, nenhum browser atualmente aceita a XSLT por completo.
>
> Você usa XSLT para manipular documentos, alterando e trabalhando com sua marcação como desejar. Uma das transformações mais comuns é de documentos XML para documentos HTML
>
> Para criar a transformação XSLT precisamos de dois documentos – documento a transformar e a folha de estilo que especifica a tranformação. Os dois são documentos XML bem formatados.

3.4.8. LINKS PODEROSOS EM XML COM XLINKS E XPOINTERS

Segundo Holzner (2001, p.21),

> [...] é dificil imaginar a WWW sem hiperlinks; logicamente, os documentos HTML são ótimos em permitir que você ligue uma página à outra. E a XML? Em XML, você utiliza Xlinks e Xpointers.
>
> Xlinks permitem que qualquer elemento se torne um link e não apenas um único elemento como o elemento <A> da HTML. Isso é bom porque a XML não possui um elemento

> <A> embutido. Em XML, como você define seus próprios elementos, faz sentido que você possa definir quais deles representam links para outros documentos.
>
> Xlinks são mais poderosos do que simples hyperlinks. Xlinks podem ser bidirecionais, permitindo que o usuário retorne após seguir um link. Eles podem ainda ser multidirecionais - na verdade, podem ser sofisticados o suficiente para apontar para o site de espelho mais próximo do qual um recurso pode ser apanhado.

As especificações XPointer e XLink, estão atualmente em desenvolvimento, introduzindo um modelo de ligação padrão para o XML.

Segundo Kade (1999, p.18),

> Um dos principais objetivos no desenvolvimento de XML foi manter e estender os mecanismos de hipertexto, tão bem sucedidos em HTML. Entretanto, em HTML, as marcações que identificavam links eram sempre do mesmo tipo, o que acontece em XML, uma vez que o usuário pode criar suas próprias marcações. Assim, tornou-se necessário criar algum tipo de especificação para indicar que determinado elemento é um hyperlink.

DeRose et al (2001) dizem sobre a especificação Xlink:

> Esta especificação define a XML Linking Language (Xlink), que permite que elementos sejam inseridos nos documentos XML para criar e descrever links entre recursos. Ela utiliza a sintaxe XML para criar estruturas que possam descrever os hyperlinks unidirecionais simples da HTML de hoje, além de links mais sofisticados. (tradução nossa)

Em XLink, um link expressa um relacionamento entre recursos. Um recurso é qualquer local (um elemento, o seu conteúdo, ou uma parte do seu conteúdo, por exemplo) que é endereçável em um link. A natureza exata do relacionamento entre os recursos depende da aplicação que processa o link e da informação semântica fornecida. Alguns destaques do XLink são:

- O XLink possibilita o controle sobre a semântica do link.

- O XLink introduz Links Extendidos que podem envolver mais de dois recursos.

Visto que a XML não tem um conjunto fixo de elementos, o nome do elemento de ligação não pode ser usado para localizar links. Ao invés disso, os processadores XML identificam os links pelo reconhecimento do atributo <xml:link>.

Os links Xlink podem ser um tanto simples, similares aos links da tag <A> do HTML que permitem algumas configurações, mas todas elas implícitas da tag <A>, ou ainda muito mais complexos, que é o verdadeiro significado da criação do Xlink.

Holzner (2001, p.574) fala sobre os links HTML:

> Há muita funcionalidade aqui, mas toda ela se baseia no elemento <A> e no tipo mais simples de hyperlink – um que espera para ser clicado e depois navega para um novo documento ou local de documento.

O autor diz, ainda, que

> Os relacionamentos entre documentos podem ser muito mais complexos do que isso. Por exemplo, você poderia querer fazer um ou mais dos seguintes:
>
> - Configurar um link para apontar para dez sites espelho de um site principal e deixar que o browser selecione o que estiver mais próximo.
>
> - Vincular a um conjunto inteiro de documentos (incluindo subconjuntos), no qual o browser deverá procurar o recurso que você deseja.
>
> - Configurar uma série de caminhos que permitam ao usuário navegar entre um conjunto de documentos em várias direções, mas não em outras.
>
> Xlinks não estão restritos a qualquer elemento, como o elemento <A>, o que significa dizer que Xlinks podem, nem sempre, aparecer nos seus documentos no texto tradicional azul e sublinhado (embora naturalmente, também possam, se você quiser que apareçam dessa maneira). Poder transformar qualquer elemento em um Xlink é ótimo, pois você pode criar elementos que sempre são links para outros recursos.

Os links em XML podem ser simples ou estendidos; nos links simples, um link XML aponta para um único recurso. Isto é especificado como o valor de seu atributo HREF, que é obrigatório. Eles podem ter outros atributos também.

Um link simples é uma via de mão única para seu destino, e normalmente é definido da seguinte maneira:

```
<link xml:link="simple" href="locator">Texto do Link</link>
```

EXEMPLO 32 – LINK SIMPLES

Um link simples (link:”simple”) identifica um link entre dois recursos, um dos quais é o próprio conteúdo do elemento do link. Um localizador (href=”locator”) identifica o outro recurso, pode ser um URL, uma consulta ou um ponteiro.

Apesar de úteis as extensões XML para links simples, elas têm as mesmas limitações dos links in-line, que são aqueles links cujos recursos estão realmente contidos no elemento de vínculo. Os links entendidos, porém, permitem criar links de mão dupla ou vincular mais de um alvo.

Segundo Kade (1999, p.18),

> [...] os links entendidos podem conectar qualquer número de recursos, além de que a definição dos recursos podem ser armazenadas fora do documento XML, o que permite que os links sejam atualizados com maior facilidade, caso um recurso mude de lugar.

Links Estendidos - permitem expressar relacionamentos entre mais de dois recursos:

```
<CITACAO xml:type=”extended" xlink:role="annotation">
    <locator xml:link="locator" href="doc1.xml">Autor</locator>
    <locator xml:link="locator" href="doc2.xml">Autor2</locator>
    <locator xml:link="locator" href="doc3.xml">Anotações</locator>
```

```
        <locator xml:link="locator" href="doc4.xml">Resenha</locator>
    </CITACAO>
```

EXEMPLO 33 – LINKS ESTENDIDOS

Segundo Holzner (2001, p.583),

> Em termos gerais, portanto, um link estendido é composto de conexões entre um conjunto de recursos. Tais recursos podem ser locais, o que significa que realmente fazem parte do elemento de link estendido, ou remotos, o que significa que não fazem parte do elemento de link estendido (mas isso não significa que precisa estar em outro documento). Se um link não contém quaisquer recursos locais, é denominado fora-de-linha.
>
> Assim, como uma aplicação utiliza um link estendido? Isso fica totalmente a critério da aplicação.

Ainda hoje não há softwares disponíveis que implementem os links estendidos, mas, com certeza, esse recurso deverá ser muito utilizado na construção de ferramentas de busca com agentes inteligentes.

Os XPointers oferecem a sintaxe que lhes permite localizar um recurso através da árvore de elementos do documento que contém o recurso, ou seja, fornece um mecanismo para identificar objetos em um documento pelo seu contexto e não somente por um identificador único.

Conforme Ramalho (2000),

> O Xpointer vem complementar o Xlink. Normalmente, quando temos uma ligação de uma página à outra e se indica ao browser que se quer seguir essa ligação, o browser carrega a nova página integralmente. A idéia por detrás do Xpointer é otimizar esta funcionalidade. Para isso, disponibiliza uma pequena linguagem de query que permite selecionar qual a parte da nova página se quer ver.

Através da Xpointer é possível definir links sem a necessidade de dizer onde está o alvo. Os alvos podem ser a numeração de um parágrafo, a primeira ou qualquer outra ocorrência de uma palavra ou de um

elemento, sendo possível especificar o intervalo do documento onde deverá ser realizada a consulta para fazer o link.

Através do Xpointer deixamos de ter links fixos e passamos a ter links definidos através de variáveis dentro do próprio documento.

A linguagem XML torna-se, assim, um ótimo instrumento para a organização da informação digital, pois permite que a informação seja descrita de uma maneira que possibilite a interoperabilidade entre diferentes meios de acesso, uma vez que, juntamente com a informação, se descreve o significado da mesma.

Santarem Segundo et al. (2003, p.4) observam que

> [...] do ponto de vista do desenvolvimento de bases de informações, a XML está aproximando alguns grupos bastante distintos em um novo conflito: desenvolvedores de documentos tentando entender questões relacionadas a sistemas gerenciadores de banco de dados; analistas de bancos de dados confusos porque o modelo relacional não atende mais as suas necessidades; e profissionais da Web que têm de lidar com transformações baseadas em esquemas e regras. A chave para resolver esse conflito é entender as estruturas semânticas diferentes que estão por trás dos padrões da XML e como modelar essa semântica para atingir os objetivos de cada um desses grupos.

Este ponto é abordado no Capítulo 4 – A recuperação da informação na web: das ferramentas de busca às potencialidades do delineamento de uma “web semântica” .

Neste capítulo discutimos os métodos de armazenamento de informação. A seguir veremos como recuperar as informações que são armazenadas na Web.

CAPÍTULO 4
A RECUPERAÇÃO DA INFORMAÇÃO NA WEB: DAS FERRAMENTAS DE BUSCA ÀS POTENCIALIDADES DO DELINEAMENTO DE UMA "WEB SEMÂNTICA"

4 A RECUPERAÇÃO DA INFORMAÇÃO NA WEB: DAS FERRAMENTAS DE BUSCA ÀS POTENCIALIDADES DO DELINEAMENTO DE UMA "WEB SEMÂNTICA"

As possibilidades da Web é um assunto que estimula inúmeras pesquisas, suscitando discussões sobre as metodologias utilizadas para estes estudos, que nem sempre são claras e oferecem um grau de incerteza muito grande quanto à medida de seu tamanho.

Recentemente, novas pesquisas estão sendo publicadas sobre a questão da "Internet invisível ou *Deep Web*", a parte da Internet que não é acessível através dos mecanismos de busca. Segundo Bergman (2001), "Informação disponível na Web invisível é 400 a 550 vezes maior do que a comumente definida na World Wide Web" (tradução nossa). Isso se deve ao fato de existirem inúmeros bancos de dados mantidos por agências governamentais, universidades e companhias privadas.

Independente do tamanho estimado, a Internet pode ser vista como uma grande massa de informações. Localizar informações sobre um determinado tema na Web, muitas vezes, pode ser um trabalho muito mais complicado do que imaginamos.

Para encontrar a informação desejada existe, hoje, uma coleção mais do que centenária de mecanismos de localização nas páginas Web, que chamamos de ferramentas de busca. Essas ferramentas vêm acompanhando a Internet há algum tempo, desde antes da popularização da Web, como é o caso do Archie para busca em FTP e do Verônica e Jughead para pesquisa em Gopher.

4.1 FORMAS DE LOCALIZAÇÃO, INDEXAÇÃO E DESCRIÇÃO COMO SUPORTE À RECUPERAÇÃO DA INFORMAÇÃO

As ferramentas de busca são instrumentos de localização de informação na WEB através de um dado informado pelo usuário.

Segundo a enciclopédia digital Webopedia (2003), os mecanismos de busca (“search engines”) são:

> Programas que pesquisam em documentos por palavras-chave especificadas e recuperam uma lista de documentos onde as palavras-chave foram encontradas. Embora ‘mecanismos de busca’ sejam uma classe geral de programas, o termo é freqüentemente usado para especificamente descrever sistemas como AltaVista e Excite que permitem aos usuários pesquisar por documentos na World Wide Web e newsgroups USENET.
>
> Tipicamente, um mecanismo de busca trabalha enviando um robô (*spider*) para buscar o maior volume de documentos possível. Outro programa, chamado indexador, lê esses documentos e cria um índice baseado nas palavras contidas em cada documento. Cada mecanismo de busca usa um algoritmo próprio para criar seu índice de tal modo que, em condições ideais, só resultados significativos sejam recuperados para cada busca. (tradução nossa)

Os mecanismos de busca existentes se diferenciam em vários aspectos, tais como: formas de localização, descrição, indexação e recuperação das informações, diferentes recursos de busca disponíveis, forma de apresentação das informações para o usuário.

Geralmente as ferramentas de busca são classificadas em diretórios, motores de busca ou índices e metamotores ou metapesquisadores.

Os diretórios e motores/índices possuem uma base de dados contendo representações (metadados) das páginas que indexam. Os metadados utilizados variam em cada serviço, incluindo desde o endereço do site (URL) até o texto integral ou etiquetas de marcação e posicionais da página (*tags*).

A diferença básica entre esses dois tipos de serviço está na forma em que a base de dados é produzida.

Os diretórios foram os primeiros a aparecer e, entre as características principais, ocorre a categorização das informações recuperadas e apresentadas, fazendo com que a recuperação e o resultado sejam apresentados em categorias e subcategorias, de acordo com o assunto abordado. O trabalho de indexação é feito, em muitos casos, por edição humana. A divisão por assuntos é elaborada utilizando-se uma estrutura hierárquica, em que cada site é indexado em um ou mais assuntos sob uma estrutura de árvore.

A maneira como a estrutura hierárquica de assuntos é dividida está diferenciada nas ferramentas, sendo que algumas utilizam inclusive linguagens e sistemas de representação mundialmente conhecidas como, por exemplo, o a lista de cabeçalhos da Biblioteca do Congresso Americano (Library of Congress Classification), Sistema de Classificação Decimal Universal (CDU) ou ainda o Sistema de Classificação Decimal de Dewey (CDD).

Como exemplos das inúmeras ferramentas classificadas como diretório, temos o Yahoo!, Cadê?, Zoom, Busca Brasil, LookSmart e Open Directory.

Para um entendimento de algumas características de uma ferramenta do tipo diretório, descrevemos a seguir o Yahoo! Brasil.

O Yahoo! é uma ferramenta criada em 1994 por dois pesquisadores norte-americanos chamados: David Filo e Jerry Yang.

Para se ter um site indexado no Yahoo! é necessário que seja feito um cadastro do site, ou seja, o usuário que é o proprietário do site faz a indicação para o Yahoo! do seu site e aguarda uma avaliação da própria Empresa, para que este novo site possa pertencer ao seu banco de dados.

O Yahoo! denomina “Surfistas do Yahoo!” o grupo de pessoas que trabalha indexando os novos sites e classificando os mesmos em categorias e sub-categorias.

De acordo com o documento “Yahoo! Brasil: perguntas mais freqüentes” (2003), disponível no site do Yahoo,

> Todos os sites sugeridos são visitados e avaliados pelos Surfistas do Yahoo! que decidem onde colocá-los. Isso garante que o Yahoo! seja organizado da melhor forma possível, tornando o diretório fácil de usar, intuitivo, útil e claro para todos.

O grande dificultador de se ter uma ferramenta de busca que indexa as informações manualmente é o tempo que uma informação demora para ser localizada e indexada. O limite de informações, que os seres humanos têm capacidade de indexar diariamente, faz com que o banco de dados desse tipo de ferramenta seja menor do que o banco de dados de uma ferramenta do tipo índice, apesar de mais preciso.

As ferramentas de busca, do tipo motores de busca ou índices, possuem robôs de busca (tanto conhecidos como: *spiders* - aranhas, *wanderers* – agentes viajantes ou *crawlers* - rastejadores) que varrem os sites da Internet, em especial da World Wide Web, seguindo os links e descrevendo e indexando automaticamente a informação coletada pelos robôs.

Exemplos deste tipo de ferramenta de busca são Google, AltaVista, TodoBR, Radar UOL, Teoma, HotBot, Northern Light, e AllTheWeb.

As ferramentas que trabalham com robôs de busca normalmente se interessam em selecionar a informação, alimentar os bancos de dados com a maior quantidade possível de informação. Esse trabalho, efetuado pelos robôs, inicia na localização de uma URL, passa pela descrição automática das informações contidas no site e termina no registro das

informações descritivas e temáticas na imensa base de dados, sem nenhuma interferência humana.

As ferramentas de busca que utilizam robôs não trabalham exatamente da mesma maneira, desde o armazenamento até o resultado, quando na recuperação de informações pelos usuários as técnicas são diferentes.

Neste trabalho, abordamos brevemente a ferramenta Google, um dos mais populares motores/índices da Web.

A ferramenta de busca Google surgiu na Universidade de Stanford (Stanford University - EUA), através de um projeto de Sergey Brin e Lawrence Page em 1998, quando fundaram então a empresa Google.

O método de pesquisa funciona com uma lógica de descrição e indexação semelhante à do Altavista e do AllTheWeb, utilizando-se de robôs, denominados *crawlers*, que estão navegando continuamente na Web, localizando, descrevendo, indexando e registrando as informações descritivas e temáticas em seus bancos de dados. A diferença mais significativa é o método para a ordenação dos resultados das buscas para o usuário final.

O Google desenvolveu um método chamado de *PageRank* que define uma lógica interna para inserir um site em seu banco de dados, assim como define a posição em que o site vai aparecer para os usuários. Essa metodologia consiste em citações, ou melhor, links das outras páginas para esta. O que o sistema faz, é calcular um valor que a página tem com base em quantos links são feitos para aquela página. Também existe a lógica de valoração de ranking, quando uma página com o PageRank muito elevado faz um link para a página que está sendo analisada. Se imaginarmos uma página em que o Servidor UOL (Universo On Line) faz uma referência diretamente, podemos inferir que essa página

deve ser importante e, como tal, deve ser também muito valorizada. Ou seja, é a lógica que indica quanto valorizada é uma página em função de ser citada por muitos Web sites ou por Web sites altamente conhecidos e valorizados.

A *PageRank* é definida da seguinte forma pelo próprio Google, em seu documento Porque usar o Google (2003):

> A classificação das páginas (PageRank) confia na natureza excepcionalmente democrática da Web, usando sua vasta estrutura de links como um indicador do valor de uma página individual. Essencialmente, o Google interpreta um link da página A para a página B como um voto da página A para a página B. Mas o Google olha além do volume de votos, ou links, que uma página recebe; analisa também a página que dá o voto. Os votos dados por páginas importantes pesam mais e ajudam a tornar outras páginas importantes.
>
> Sites importantes, de alta qualidade recebem uma nota de avaliação maior, que o Google grava a cada busca feita. Naturalmente, uma página importante não significa nada se não combinar com a sua busca. Assim, o Google combina os resultados de alta qualidade com a busca que você está realizando, para que o resultado seja o mais relevante possível. O Google pesquisa quantas vezes a palavra procurada aparece nas páginas e examina todo o aspecto delas (e conteúdo das páginas ligadas a ela) para determinar o melhor resultado para a sua busca.

O Google indexa hoje, segundo informações contidas em seus sites, cerca de 3,3 bilhões de páginas na Internet, valor próximo do AllTheWeb com 3,15 bilhões.

Para Cendón (2001, p.42),

> A maioria dos motores de busca indexa, ou seja, inclui em seu índice, cada palavra do texto completo, apenas o URL, as palavras que ocorrem com frequência ou palavras e frases mais importantes contidas no título ou nos cabeçalhos e na primeiras linhas, por exemplo. Alguns motores indexam outros termos que não fazem parte do texto visível, mas que contém informações importantes e úteis. Exemplos deste tipo de texto são os textos ıncluídos nos metatags para classificação, descrição, palavras-chave e texto ALT da tag

> <IMAGE>, ou seja, texto associado com imagens. Os metatags de classificação forncecem uma palavra-chave que define o conteúdo da página. Os de descrição retornam à descrição da página feita pelo seu autor no lugar do resumo que o robô criaria automaticamente. Os de palavras-chave fornecem palavras-chave designadas pelo autor para descrever seu conteúdo ou assunto. Por exemplo, no metatag <META name="keyword" content="Brasil, informação para negócios">., as palavras Brasil e informação para negócios podem não fazer parte do texto visível da página, entretanto foram indicadas pelo seu autor como indicadores do assunto sobre os quais a página versa.

Algumas ferramentas de busca também têm um serviço que um cliente, por meio de contrato e pagamento, mantém seu site na lista dos resultados, podendo, inclusive em alguns casos, dependendo do valor pago, inserir seu site entre os primeiros que aparecem em um resultado de busca, quando a informação solicitada corresponder ao site contratado. O Yahoo!, AltaVista e AllTheWeb são sites que oferecem esse tipo de serviço, enquanto que o Google não oferece esse tipo de transação comercial.

A freqüência de atualização dos dados das ferramentas de busca varia de acordo com seus objetivos. Na maioria delas, existem robôs específicos fazendo a varredura de links que podem estar desativados e devem ser retirados do banco de dados. Normalmente, a atualização das páginas nos principais sites de busca, inclusive nos descritos anteriormente, tem freqüência de atualização entre 2 a 4 semanas.

Além das classificações de diretório e índices, temos como terceiro tipo de ferramentas os metamotores ou metapesquisadores, que são serviços que não possuem uma base de dados própria e, sim, um software que pesquisa dados de outras bases de ferramentas de busca.

O sistema de busca consiste em uma metaferramenta que envia a pesquisa para mais de uma ferramenta de busca que, na maioria das vezes, pode ser selecionada pelo usuário. Geralmente, na exibição do

resultado as duplicatas são retiradas. Exemplos desse tipo de serviço são o MetaCrawler, MetaMiner, Mamma, Kartoo e WebCrawler.

Como exemplo de um metapesquisador brasileiro, temos o MetaMiner que trabalha fazendo buscas nas ferramentas Achei, Radar UOL e LookSmart.

Além das classificações apresentadas, novos serviços de busca vêm aparecendo na Web constantemente. Entre os principais serviços, podemos destacar a busca de informações específicas por determinada área de conhecimento, dicionários, notíciários, imagens, pessoas e vídeos. Estes serviços são considerados ferramentas especiais.

4.2 RECURSOS PARA RECUPERAÇÃO DE INFORMAÇÕES

As ferramentas de busca de um modo geral oferecem ao usuário um modelo padrão, normalmente um campo onde o usuário pode digitar a informação desejada e, então, clicar em um botão para que seja iniciada a pesquisa. Essa busca padrão ou básica, denominada busca simples, é comumente utilizada. Porém, grande parte dos sites de busca também oferecem um tipo de pesquisa avançada que pode ser utilizada por usuários mais experientes.

Para os dois tipos de busca, básica ou simples e avançada, as ferramentas, geralmente, oferecem uma área de ajuda onde o usuário pode tomar conhecimento das informações sobre o funcionamento da ferramenta e as características fundamentais para elaboração de uma estratégia de busca bem elaborada, o que culmina em uma melhoria da qualidade das informações recuperadas.

Ao receber uma estratégia de busca, a ferramenta inicia um processo de recuperação em seus próprios bancos de dados, nos casos dos catálogos e índices, e uma busca nos bancos de dados das

ferramentas selecionadas, no caso dos metapesquisadores. Vale salientar que a busca não está sendo elaborada diretamente na Internet ou na Web, mas que a recuperação da informação acontece somente em registros informacionais que já tenham sido submetidos à análise de pessoas ou de *rankings* automáticos e que foram inseridos nos imensos bancos de dados das ferramentas de buscas.

A busca simples não exige que o usuário conheça lógica booleana, mas as ferramentas, internamente, têm suas metodologias de implicitar os operadores na busca do usuário. Os principais sites de busca como Yahoo!, Google, AltaVista e AllTheWeb utilizam o operador booleano E (AND) quando o usuário digita mais de uma palavra para fazer a busca.

Nas pesquisas avançadas a recuperação normalmente é mais qualitativa, pois o usuário consegue selecionar o que realmente deseja, porém, para utilizar este tipo de busca, é necessário conhecimento da ferramenta por parte do usuário. A grande parte dos sites de busca oferecem uma área de busca avançada com o intuito do usuário conseguir melhor proveito das ferramentas no momento de formular a busca.

> As Ferramentas de Busca, geralmente, trabalham com operadores booleanos, posicionais, truncamento e a combinação destes, além de recursos adicionais como a busca por linguagem natural, obrigatoriedade ou não da ocorrência do termo, diferenciação entre maiúscula e minúscula, diferenciação de acentos e caracteres especiais, além de outros que exigem conhecimento de conceitos prévios e da sintaxe da busca, como busca por data, por domínio, por URL, por título, por outros idiomas, tipo de documento. A utilização correta dos operadores e dos demais recursos interferem na qualidade das respostas obtidas com relação à sua pertinência e exaustividade [...]. (BUENO; VIDOTTI, 2000, p.13).

Além da habilidade do usuário é importante também ressaltar a capacidade dos sites em facilitar a busca das informações; para isso as ferramentas oferecem alguns recursos tais como: operadores lógicos,

proximidade entre termos, truncamento, linguagem natural, busca por tipo específico de arquivo, de domínio e de idioma.

Os operadores lógicos ou booleanos como também são chamados, são as formas mais usuais de relacionamento entre termos. Esta é uma característica presente em quase todos os mecanismos de busca, geralmente sob o rótulo de “busca avançada”. Um problema comum é que, às vezes, o relacionamento é automático ou implícito e, nem sempre, é facilitado ao usuário identificar o operador booleano que está sendo considerado quando digita apenas os termos, sem utilizar os conectores, ou seja, a operação *default*/padrão. Os principais operadores booleanos são AND (E), OR (OU) e o NOT (NÃO). Segundo o *Projeto CCN* do IBICT (1997),

> Operador lógico AND tem a função de efetuar a interseção de dois ou mais termos de busca a serem recuperados num mesmo registro. O AND tem caráter restritivo, permitindo o refinamento em um conjunto de documentos. Ao executar a pesquisa em duas chaves de busca combinadas pelo operador AND são recuperados os registros que contém a primeira e a segunda' chave simultaneamente.
>
> Exemplo: recuperar todos os registros que contenham eventos realizados em Salvador em 1992 sobre pediatria, executar a seguinte expressão booleana: pediatria and salvador and 1992.

Consta ainda no projeto,

> Operador lógico OR , tem a função de efetuar a união de dois ou mais termos de busca, permitindo recuperar documentos que tenham um ou outro dos termos utilizados. O OR amplia a recuperação de um conjunto de documentos. Ao executar a pesquisa em duas chaves de busca combinadas pelo operador OR, são recuperados os registros que contêm a primeira ou a segunda chave, ou ambas.
>
> Exemplo: recuperar em uma base todos os documentos em inglês ou francês, executar a seguinte expressão booleana: inglês or francês.

O operador lógico NOT é utilizado toda vez que se deseja fazer uma busca por uma palavra, mas sem que haja a ocorrência de outra palavra. Exemplo: se desejamos recuperar informações com a palavra direito, porém não queremos saber sobre direito constitucional, devemos elaborar a seguinte estratégia de busca: direito NOT “direito constitucional”.

Outra forma de relacionamento é através da proximidade entre termos. Em sistemas de recuperação tradicionais é comum a existência do operador NEAR (próximo), ou de operações lógicas que permitam especificar a distância máxima permitida entre dois termos de busca dentro de um registro. Esta função considera a hipótese de que quanto mais perto dois termos estejam dentro de um único texto, maior a probabilidade de estarem relacionados ao mesmo conceito.

Alguns mecanismos de busca na Web disponibilizam o recurso de proximidade, porém não é comum o uso desse operador. Segundo o documento Buscando termos perto de outros (2003), publicado no site do CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico),

> O operador de proximidade é unidirecional da esquerda para a direita. Ele recuperará apenas os registros nos quais o termo 2 ocorre em até n termos depois do termo 1. As ocorrências do termo 1 em até n termos depois do termo 2, não serão consideradas.

A chamada truncagem, truncamento ou busca por raiz é a possibilidade de busca por prefixo ou sufixo, ou ainda por um “curinga” para substituir uma letra ou conjunto de letras de uma palavra-chave, por exemplo: o uso de Brasil* pode produzir a recuperação de sites que contenham a palavra Brasil, Brasileiro ou Brasília. Este recurso não é comum nos mecanismos de busca, embora alguns o utilizem sem comunicar o usuário.

É importante observar se o servidor de busca informa ao usuário se esta é uma opção *default*, pois isto tem uma implicação direta na recuperação. Caso o usuário pesquise por “amor” e o mecanismo tenha como *default* a truncagem à direita, podem ser recuperados textos que contenham a palavra “amoroso”, por exemplo. Se o mecanismo tiver como *default* a truncagem à esquerda, podem ser obtidos resultados contendo “clamor”, por exemplo. Se tiver truncagem nos dois lados, são muitas as possibilidades de recuperação não significativa.

De forma simplificada, abordaremos a seguir as possibilidades de busca avançada, oferecidas pelos dois exemplos de ferramentas de busca: Yahoo! Brasil e Google.

O Yahoo! não trabalha com os operadores lógicos de maneira explícita em sua busca avançada, porém apresenta quatro formas distintas de busca de informações: todas estas palavras, a expressão exata, qualquer uma destas palavras ou ainda nenhuma destas palavras. O usuário pode optar por fazer uma combinação das opções desejadas, por exemplo: buscar a expressão exata “ordem e progresso” e a palavra “Brasil”. Quando o usuário escolhe a opção, ainda pode refinar a busca na opção de verificação em todo o documento ou apenas no título da página, por exemplo.

Quando o usuário escolhe uma destas opções oferecidas pelo Yahoo!, está indiretamente escolhendo o operador booleano a ser utilizado na busca. Se o usuário escolhe a opção “todas estas palavras”, o resultado deverá trazer somente os sites que tiverem todas as palavras que o usuário selecionou para serem recuperadas; neste caso está sendo utilizado o operador E. Se o usuário escolhe a opção “a expressão exata”, a ferramenta de busca vai entender que um resultado desejado deverá trazer somente páginas em que se encontre, como conteúdo, um conjunto de palavras exatamente igual ao conjunto de palavras digitadas pelo usuário, como se fosse uma busca por uma frase; se o usuário escolher

“qualquer uma destas palavras”, o Yahoo! retornará, como resultado, qualquer site que contenha pelo menos uma das palavras selecionadas pelo usuário; neste caso existe a aplicação do operador lógico OU (OR). Se o usuário preferir a última opção “nenhuma destas palavras”, o Yahoo! retornará apenas os sites em que as palavras selecionadas não existem.

O Yahoo! Brasil oferece também a possibilidade de realização da busca em um determinado domínio ou site; como opção temos a possibilidade de realizar a busca em: qualquer domínio, apenas nos domínios .BR, .COM.BR, .GOV.BR, .ORG.BR ou ainda designar o site ou domínio onde deverá ser realizada a busca.

Pode ser aplicada na busca avançada no Yahoo! Brasil a opção de filtro para conteúdo adulto: ligado ou desligado. No caso da opção ligado, o Yahoo! não permite que sejam listados no resultado sites que contenham conteúdo impróprio para menores de 18 anos.

O Yahoo! permite ainda que o usuário escolha páginas em determinado idioma ou escolha um ou mais idiomas das 35 línguas oferecidas.

A busca avançada do Yahoo! permite ainda ao usuário escolher o número de resultados por página, oferecendo as opções de 10, 15, 20, 30 e 40 resultados por página.

Na opção de pesquisa avançada o Google se parece muito com o Yahoo!, porém, podemos destacar algumas diferenças.

Assim como o Yahoo!, o Google também oferece quatro opções de pesquisa que podem ser combinadas entre elas. As opções são: com todas as palavras, com a expressão, com qualquer uma das palavras ou ainda sem as palavras, sendo que elas funcionam exatamente iguais as opções de: todas estas palavras, a expressão exata, qualquer uma destas palavras, nenhuma destas palavras respectivamente oferecidas pelo Yahoo!.

No Google, o usuário pode igualmente, selecionar o idioma em que deseja pesquisar, sendo oferecidos 35 idiomas diferentes, dos quais o usuário pode escolher apenas um ou então todos.

Temos como opção também a busca por determinados tipos de arquivo que se trata de uma inovação da busca avançada do Google; nesta opção é possível o usuário, se for de seu interesse, pesquisar apenas informações com os seguintes formatos: Adobe Acrobat (PDF), Adobe PostScript (PS), Microsoft Word (DOC), Microsoft Excel (XLS), Microsoft PowerPoint (PPT), Rich Text Format (RTF), tendo a opção ainda de escolher todos os formatos.

É oferecida também ao usuário a opção de selecionar apenas as páginas atualizadas: em qualquer data, nos últimos 3 meses, nos últimos 6 meses ou no último ano.

O Google permite selecionar as ocorrências das palavras selecionadas pelo usuário em: qualquer lugar da página, no título da página, no corpo da página, no endereço da página ou ainda em links para a página.

O Google também permite ao usuário escrever o domínio onde deverá ser realizada a busca. Neste caso, o Google oferece um campo para o usuário digitar o domínio escolhido, como exemplo: .BR, .COM.BR etc.

O Google oferece ainda uma opção de descobrir quais são os sites que fazem link para um outro determinado site na opção: encontrar páginas com link para a página e, em seguida, aparece um campo para ser digitado o endereço que se deseja verificar.

4.3 FORMAS DE APRESENTAÇÃO DAS INFORMAÇÕES

Os sites de busca citados até agora, entre eles Google, AltaVista, AllTheWeb e Yahoo!, oferecem recursos avançados dentro da área de pesquisa básica; para isso é necessário que o usuário conheça os comandos necessários para que se possa, através de uma caixa de entrada de informação, determinar a pesquisa que deseja fazer. Um exemplo disto é a pesquisa por domínio que pode ser feita diretamente na entrada de busca simples. Tomemos como base a busca exata do termo "Ciência da Informação" no site "www.marilia.unesp.br" a ser realizada no Google e no Yahoo!: como estratégia de pesquisa devemos digitar a expressão de busca da seguinte maneira: "Ciencia da Informação", site:www.marilia.unesp.br; neste caso, verificaremos que os resultados apresentarão apenas ocorrências que contenham o termo exato Ciência da Informação e que estejam dentro do site www.marilia.unesp.br.

O Google interpretou a busca exata do termo, porque na digitação o termo Ciência da Informação foi colocado entre aspas e o filtro pelo site escolhido foi o comando "site:" juntamente com a especificação do site desejado.

O resultado da busca foi apresentado pelo Google da seguinte maneira:

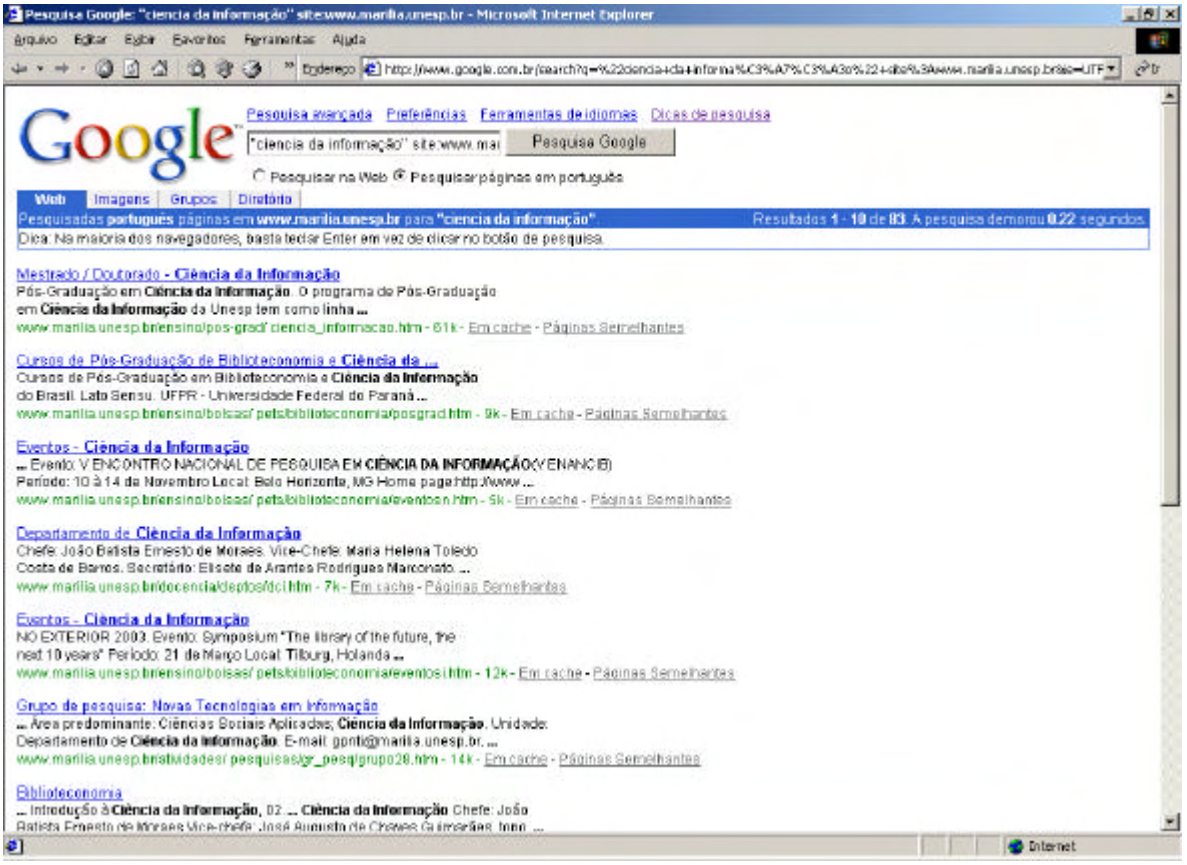

FIGURA 1 – PESQUISA NO GOOGLE

FONTE: WWW.GOOGLE.COM.BR

ACESSO EM: 15/09/2003.

O Yahoo! apresentou o seguinte resultado:

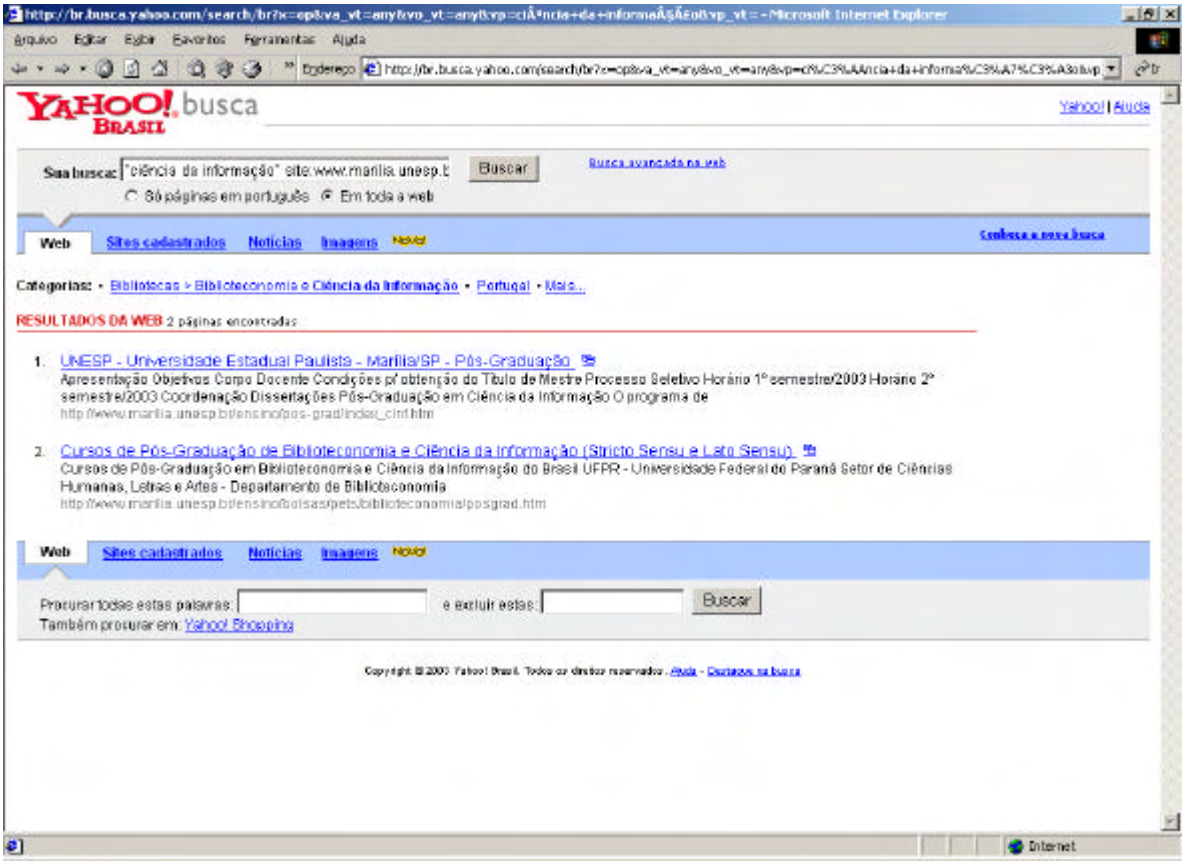

FIGURA 2 – PESQUISA NO YAHOO!

FONTE: WWW.YAHOO.COM.BR

ACESSO EM: 15/09/2003.

Quando um usuário realiza uma busca nem sempre o que se procura é listado como resultado; esse é um dos grandes problemas encontrados pelos usuários, quando necessitam de uma informação; a qualidade na recuperação da informação está sendo estudada, o desempenho de um sistema de recuperação de informação pode ser julgado pela satisfação do usuário em ter sua necessidade informacional atendida.

A necessidade de se verificar se um sistema de busca é realmente eficiente levou ao desenvolvimento de métricas para avaliação dos mesmos. No próximo tópico estaremos verificando essas métricas.

4.4 MÉTRICAS PARA AVALIAÇÃO DO GRAU DE SATISTAÇÃO DO USUÁRIO

Para se medir a satisfação do usuário alguns critérios podem ser avaliados como: Precisão, Revocação, Cobertura, Formato de Apresentação, Tempo de Resposta e Atualização do Conteúdo.

Bueno e Vidotti (2000, p.18-19), concluem que

> existe um campo de pesquisa amplo e pouco explorado para os profissionais da Biblioteconomia, com relação aos tratamentos descritivos e temáticos dos documentos disponíveis na Internet, que são catalogados em bases de dados pelas ferramentas de busca de forma automática ou manual, para que esses documentos possam ser recuperados de maneira a atender as expectativas do usuário numa relação eficaz de relevância X pertinência X tempo de busca.

Precisão é a fração das respostas retornadas que é relevante ou de interesse do usuário. Por exemplo, se três das dez primeiras respostas retornadas são relevantes, dizemos que a precisão é de 30%. Tecnicamente poderíamos dizer que é o resultado da relação entre o número de informações relevantes recuperadas (R) e o total de informações recuperadas (T), ou seja: Precisão=R/T.

No caso da precisão é muito comum falarmos em spam, que acaba influindo diretamente nos resultados. De acordo com Cendon (2001, p.46),

> Spam pode ser definido como um conjunto de métodos considerados pouco éticos para promover páginas através da repetição de palavras irrelevantes, mas muito procuradas (como, por exemplo, futebol), para que estas páginas, embora não relacionadas com a consulta, sejam localizadas por buscas comuns. Técnicas usuais de spam são o uso de texto invisível (texto escrito da mesma cor do fundo da página, e que portanto, apesar de poder ser lido pelos robôs, não é visto pelo usuário), texto escrito em letras muito pequenas, que também são difíceis de serem vistas, ou a inclusão de palavras não apropriadas nos metatags. Alguns robôs podem detectar esta repetição desnecessária de

> palavras e penalizar a página na ordenação por relevância, ou mesmo excluí-las de seu índice.

Além de spam, a cada dia surgem novas técnicas e métodos anti-éticos que vão sendo desenvolvidos com o intuito de indexar com alto grau de relevância os sites nas ferramentas de busca.

Revocação é a fração de todas as respostas relevantes que foram retornadas pela máquina de busca. Por exemplo, se há 8 respostas relevantes para uma dada consulta e a máquina de busca retornou somente duas destas dentre suas dez primeiras respostas, dizemos que a revocação é de 25% (isto é, um quarto dos documentos relevantes foram recuperados). Tecnicamente poderíamos dizer que é o resultado da relação entre o número de informações relevantes recuperadas (R) e o total de informações relevantes armazenadas (A), ou seja: Revocação=R/A

Conforme o documento, Análise comparativa de máquinas de busca para a web brasileira (2003),

> No caso da Web, focalizamos as medidas de precisão nas 10 primeiras respostas retornadas. Assim, medimos a precisão média quando observamos o primeiro documento, o segundo, e assim por diante, até o décimo. Este cálculo é feito da seguinte forma:
>
> Considere uma máquina de busca à qual são submetidas 10 consultas de teste. Se observamos a primeira (ou melhor) resposta para cada uma das 10 consultas, notamos que ela pode ser relevante ou não relevante. Assim, para cada consulta, a precisão quando observamos somente a primeira resposta é 0% ou 100%. Para as 10 consultas, a precisão média para a primeira resposta é calculada através da média dos 10 valores de precisão correspondentes. Se observamos as duas primeiras respostas para cada uma das 10 consultas, distinguimos três situações possíveis: as duas respostas são relevantes, uma resposta é relevante e a outra não é, as duas respostas não são relevantes. Assim, para cada consulta, a precisão quando observamos somente as duas primeiras respostas é 0%, 50%, ou 100%. Para as 10 consultas, a precisão média para as duas primeiras

> respostas é calculada através da média dos valores de precisão correspondentes. De modo análogo, podemos repetir este procedimento para obter a precisão média para as três primeiras respostas, para as quatro primeiras respostas, e assim por diante.

Tratando do terceiro parâmetro, a cobertura, presume-se que esta seria a solicitação mais importante de usuários: que o serviço ofereça todo o volume de informações disponíveis sobre determinado assunto. Porém, de acordo com a necessidade do usuário de alta precisão ou alta revocação, como exposto anteriormente, a cobertura deixa de ser um item tão significativo.

A avaliação da cobertura deve ser feita considerando dois aspectos: abrangência e escopo.

A abrangência diz respeito ao volume de informações que o mecanismo de busca registra proporcionalmente ao tamanho da Web. Esta é uma medida difícil de ser avaliada, visto que as estatísticas refletem informações oferecidas pelos próprios mecanismos de busca e porque o tamanho da Web também é imensurável.

Em relação ao escopo, deve-se avaliar que tipo e que formato de recursos o mecanismo de busca inclui. Alguns serviços indexam, além de páginas html, documentos em arquivos formatados (PDF, DOC, XLS por exemplo), imagens, mensagens de grupos ou listas de discussão.

É importante avaliar se o mecanismo de busca descreve e indexa o texto completo das páginas e quais os metadados que são armazenados em seu banco de dados. Em geral, são registrados a data do documento e o tamanho em bytes; há serviços que incluem o idioma das páginas.

O formato de apresentação é um item que tem apresentado pouca diferença entre as principais ferramentas de busca; a cada dia as ferramentas procuram aumentar o conteúdo de informação de cada resposta procurando dar o máximo de informação sobre o resultado

apresentado antes do usuário entrar no site listado. O Google, por exemplo, oferece os seguintes itens:

- Barra de Estatísticas: essa linha descreve a consulta e indica o número de resultados retornados, assim como a quantidade de tempo necessária para completar a consulta;
- Categorias: se os termos da sua procura também aparecem no diretório, estas categorias sugeridas podem ajudar a encontrar mais informações relacionadas à consulta;
- Título da página: a primeira linha do resultado é o título da página Web encontrada. Ocasionalmente, ao invés do título, encontrará um URL, o que significa que essa página não tem título ou o Google não indexou todo o conteúdo dessa página. Se o texto associado com esses links coincide com o texto que introduziu para a pesquisa, a página aparecerá como resultado, mesmo que todo o seu texto não tenha sido indexado.
- Texto abaixo do título: este texto é um resumo da página-resultado com os seus termos de pesquisa em negrito. Estes resumos permitem prever o contexto cujos termos de pesquisa aparecem na página.
- Descrição: se o texto de pesquisa está listado no diretório Web, é exibida a descrição pelo autor do diretório aberto.
- Categoria: se um site encontrado pelo seu texto de pesquisa está listado no diretório Web, a categoria da qual faz parte é mostrada por baixo da sua descrição.
- URL do Resultado: é o endereço Web do resultado.

- Tamanho: é o tamanho da parcela do texto da página Web encontrada. É omitido para sites que ainda não foram indexados pelo Google.
- Cache: fazer click no link em cache irá permitir visualizar o conteúdo da página no momento em que foi cadastrada. Se, por alguma razão, o link do site não o leva à página atual, pode ser visualizada a versão em cache e encontrar a informação de que necessita. Os termos de pesquisa são realçados na versão em cache.
- Páginas similares: quando seleciona o link de Páginas Similares num resultado concreto, o Google irá automaticamente procurar a Web por páginas relacionadas com esse resultado.
- Resultado recortado: quando Google acha múltiplos resultados para um mesmo site, o resultado mais relevante é listado primeiramente com as outras páginas relevantes desse mesmo local, recortado abaixo dele.

Nas outras ferramentas de busca já citadas verificamos que alguns itens são diferentes: porém, de uma maneira geral, as ferramentas têm procurado dar a maior quantidade de informação possível ao usuário.

O tempo de resposta é um critério que também influi na satisfação do usuário, porém, existem ainda grandes disparidades sobre as pesquisa que avaliam o tempo de resposta dos sites, pois isto envolve, além da velocidade de conexão do usuário que ainda é muito diferente nos mais diversos locais de acesso à Internet no Brasil, a quantidade de resultado oferecido pela ferramenta. Baseados em uma busca realizada no mesmo local, utilizando-se o mesmo computador e com a mesma velocidade de acesso, os metapesquisadores acabam não tendo a mesma performance que os catálogos e índices, visto que os metapesquisadores

têm o trabalho de submeter a pesquisa do usuário aos bancos de dados de outras ferramentas de busca e, antes de apresentá-los como resultado, devem aplicar um filtro para evitar que os resultados se repitam muito. Nos testes realizados foi verificado que, com o mesmo tipo de pesquisa, os metapesquisadores, na sua maioria, utilizam o dobro do tempo em relação às ferramentas que possuem bancos de dados próprios.

Um detalhe importante, quando tratamos de tempo de resposta e que agrada muito aos usuários, é que o mesmo saiba o que está acontecendo quando submete uma pesquisa, ou seja, que apareça uma figura ou um texto informando que a busca esta sendo realizada e que em breve será apresentado o resultado.

A atualização do conteúdo nos sites de busca é, talvez, um dos critérios mais importantes quando tratamos de recuperação da informação, porque a ferramenta vai perdendo a credibilidade junto ao usuário quando apresenta em seus resultados links quebrados/nulos, que são links para onde não existe mais a página, e também resultados com sites que não são atualizados há muito tempo.

Conforme abordado, as principais ferramentas de busca procuram atualizar constantemente seus bancos de dados, tentando evitar ao máximo passar informação desatualizada para o usuário.

De um modo geral as ferramentas de busca têm procurado técnicas para proporcionar os resultados da melhor maneira possível aos usuários: porém, algumas limitações continuam a fazer com que os resultados, na maioria das vezes, não atendam da melhor maneira o desejo dos usuários. Os principais empecilhos ainda são a quantidade de informação disponível na rede, assim como a velocidade com que as informações aparecem e desaparecem e a pouca padronização da publicação de documentos na Internet. As novas tecnologias estão propondo um novo formato para a Web, procurando organizar as informações.

Em seguida estaremos abordando a Web Semântica, como uma das promessas para dar significado às informações dispostas na Web. Talvez este seja o caminho que possibilitará às ferramentas de busca melhorar a qualidade dos resultados apresentados, dando maior precisão e confiabilidade, além de economia de tempo aos usuários.

4.5 WEB SEMÂNTICA: PADRÕES PARA ORGANIZAÇÃO DA INFORMAÇÃO DIGITAL

Durante este trabalho de pesquisa verificamos que a tecnologia passou a ser fundamental em todos os segmentos. O aparecimento de máquinas e computadores, aliado às pesquisas que multiplicam o poder de processamento e armazenamento de informações, cada vez a um custo mais baixo, fez do último século o chamado “Século da Informação”.

O uso do computador e suas características de compartilhamento de informações através de uma rede passaram a ser ponto determinante para o crescimento de algumas áreas do conhecimento, possibilitando novas formas de comunicação e alterando a relação espaço-tempo na busca de informações. A troca de informações entre grupos de pesquisadores, negociações entre empresas, atendimento ao cliente ou mesmo o uso como forma de entretenimento já fazem parte do cotidiano dos usuários dessa rede.

O uso da tecnologia tem realizado muitos feitos, porém, o uso das tecnologias de informação e comunicação não têm atendido adequadamente as exigências do ser humano, quando é necessário a tomada de decisão. Essa evolução nos permitiu acelerar o processamento das informações, assim como facilitar a socialização delas mas a partir do momento que temos a informação, continuamos a ter que tomar decisões de acordo com a nossa própria estrutura cognitiva.

Os estudos sobre inteligência artificial e redes neurais têm procurado inserir nas máquinas o processo de tomada de decisões, inclusive com uso de GPS e aplicações na área médica.

A evolução tecnológica que acelera o processamento e a troca de informações fizeram com que novas tecnologias fossem descobertas, criadas e recriadas em curto período de tempo, fazendo com que pessoas e instituições fossem desenvolvendo suas próprias maneiras de organizar e estruturar seus dados e informações, dando início principalmente à construção de bancos de dados e grandes sistemas de informação, todos independentes e, conseqüentemente, sem possibilidade de troca de informações.

Como vimos no capítulo 3, na Internet não foi diferente, mesmo tendo um ambiente interligado, as informações, são disponibilizadas utilizando a estrutura definida por quem cria a informação, não seguindo nenhum padrão de organização ou regras de descrição, mesmo porque estas não existem. Hoje em dia estão sendo recomendadas algumas técnicas para publicação de informação na Web, mas estas técnicas não tratam de regras e acabam ficando confinadas aos estudos, pesquisas e laboratórios acadêmicos.

Como a Internet é, sem sombra de dúvida, hoje o meio de comunicação que permite maior interatividade entre as pessoas na troca de informações e o seu crescimento não permite calcular precisamente a quantidade de informações ou recursos disponíveis, a abundância de recursos/documentos digitais exacerba os limites da compreensão humana.

As novas tecnologias da informação possibilitam o acesso a todo tipo de informação, em qualquer lugar e a qualquer momento. Na Internet, o acesso à informação é facilitado pelas características da rede que a diferencia dos demais meios de comunicação, uma vez que esta possui uma grande capacidade de armazenamento, facilidade de

manipulação de seu conteúdo informacional e de seus serviços e recursos, possibilidade de pesquisa e a transmissão instantânea da informação para o usuário.

Diante disso, temos o seguinte quadro: muita informação disponível, e aumentando cada vez mais, distribuídas em bancos de dados relacionais e páginas da Internet com textos, imagens, sons e vídeos, mas que ainda não podem ser recuperadas de maneira eficiente; em muitos casos, uma mesma informação não pode ser reaproveitada em situações ou necessidades diferentes, uma vez que não existe a interoperabilidade entre os dados, ou seja, um documento preparado para ser interpretado por um software, não pode ser interpretado por outros, na mesma ou em outra plataforma computacional.

Segundo Marcondes e Sayão (2001, p.26),

> A enorme quantidade de informação armazenada e disponibilizada via Internet torna cada vez mais crítico o problema da identificação de informação relevante, assim chamada information discovery. Diferentes estratégias para fazer frente à explosão informacional trazida pela Internet podem hoje ser divisadas, como os mecanismos de busca gerais (AltaVista, Excite, Lycos, Infoseek, Yahoo e outros), os localizadores de informações especializados, como o GILS (http://www.usgs.gov/gils/) ou portais temáticos como o SIGNPOST (http://www.signpost.org) americano, o OMNI (http://www.omni.ac.uk) e o SOSIG (http://www.sosig.ac.uk) ingleses, o PROSSIGA – Comunicação e Informação para a Pesquisa – (http://www.prossiga.br) ou LIS – Localizador de informações em Saúde – (http://www.bireme.br) no Brasil. Ambas as alternativas, os mecanismos de busca gerais e os portais temáticos oferecem soluções parciais para a localização de informações na Internet, principalmente as de interesse para C&T.

Os problemas dos mecanismos de busca são os mesmos; Marcondes e Sayão (2001, p.26) afirmam, ainda, que

> Entre os principais, pode-se citar os seguintes: baixa qualidade da indexação, por ser feita automaticamente, que resulta em grande quantidade de informações recuperadas,

> a maioria sem relevância (em termos de recuperação de informação, oferecem alta revocação, mas baixa precisão); cobertura parcial da Internet; as ferramentas de busca não são especializadas; indexam páginas HTML isoladas e não recursos; além disto, grande quantidade de informações disponíveis na Internet estão sob a forma de registros contidos em bases de dados, que ficam assim "escondidas"; estes registros são acessados somente por meio das interfaces destas bases de dados, o que pressupõe uma interação entre um usuário humano com a base de dados e, portanto, ficam inacessíveis aos programas robôs.

A facilidade de interpretação que o ser humano tem em distinguir uma palavra em um determinado contexto não é encontrada nos computadores e nos robôs de busca, não permitindo, assim, que os mesmos consigam entender o conteúdo significativo de uma página Web antes de descrevê-la e informá-la como resultado a um usuário.

Santarem Segundo e Vidotti (2003, p.3), enfocam que,

> Os computadores trabalham com processamento lógico, mas não são capazes de fazer associações de significados, diferentemente da mente humana que é capaz de juntar partes de informações dispersas e de estabelecer um novo contexto, identificando o significado das informações dispostas e assimilando um novo conhecimento.

A questão das dificuldades de localização, descrição e recuperação de informações conduziram à criação de um novo projeto encabeçado por Tim Berners-Lee, desenvolvedor do sistema de códigos HTML. Berners-Lee criou também o sistema de endereços que permite localizar cada página da rede, as regras para interligação desses documentos em computadores ligados à Internet e ainda o browser para navegação na rede.

Segundo Berners-Lee, Lassila, Hendler (2001), um caminho para a solução da qualidade na recuperação dos dados que permita ao usuário resultados mais precisos parece ser a criação da Web Semântica, um projeto que visa dispor nos sites tanto informações descritivas e temáticas

para os usuários, como informações que possam ser processadas e identificadas pelos computadores automaticamente. Assim, seria uma forma de disponibilizar informações para as máquinas/softwares juntamente com as informações para os usuários.

4.6 DELINEAMENTO DA WEB SEMÂNTICA, UMA NOVA PROPOSTA PARA A WEB

Segundo Berners-Lee, Lassila, Hendler (2001),

> A Web Semântica trará uma estrutura ao significado da página Web, criando um ambiente propício para que os agentes de busca possam realizar tarefas sofisticadas e entregá-las ao usuário. (tradução nossa)

A Web Semântica poderá ser, não uma nova Web, e sim a extensão da Web atual, que permitirá que computadores e pessoas possam trabalhar cooperativamente no processo de descrição, armazenamento e recuperação de informações digitais.

O desafio da Web Semântica é prover uma linguagem capaz de expressar ao mesmo tempo dados e regras, de forma a possibilitar a dedução de novos dados e regras a partir de qualquer sistema de representação de conhecimento a ser importado ou exportado na Web.

Conforme apontam Silva e Lima (2002),

> Os computadores são úteis para organização e processamento lógicos, mas não são capazes de estabelecer associações de significado. Um computador tipicamente mantém as informações em hierarquias rígidas, enquanto a mente humana tem a habilidade especial de ligar pequenas unidades de informação de forma randômica. Com base nesta constatação, a segunda geração da World Wide Web (WWW, Web), cunhada como Web Semântica [Berners-Lee, 1999], envolve o arranjo das idéias e de suas associações de forma não restrita. Assim, um computador poderia representar associações entre coisas que poderiam parecer não relacionadas, mas que de fato, compartilham algum relacionamento. Tim Berners-Lee, criador da Web, via nesta

> a necessidade de uma evolução, até que ela tenha o poder de fazer com que as informações possuam formato tal que as máquinas venham a fazer associações entre informações que se relacionam [Berners-Lee, 2001]. Quando isso ocorrer de fato, terá sido implementada a Web Semântica.

O projeto da Web Semântica tem assim, como ponto fundamental, a criação de uma nova estrutura de armazenamento de dados. O ponto principal está na separação da apresentação do conteúdo e do conteúdo da estrutura, tratando as unidades atômicas de uma informação como componentes independentes.

Essa separação permitirá uma recuperação da informação de várias maneiras, independente de como seja esta busca, bastando que se conheça a estrutura dos dados. Isto resolverá o problema da utilização de uma mesma informação em vários sistemas. Parece-nos que esse mecanismo de troca de informações, além de solucionar esse problema, permitirá o desenvolvimento de novas pesquisas em favor da interação entre vários sistemas e dados em todo o mundo.

Acrescentar regras de semântica à Web irá mudar a sua natureza de maneira radical, transformando um meio que apenas exibe informação em um meio em que a informação é interpretada, trocada e processada por softwares. Mecanismos de busca semânticos serão capazes de reunir informações de bancos de dados dispersos, processá-las e deduzir novos conteúdos automaticamente. Programas que não estavam projetados para compartilhar informação começarão a fazer isso e serão compatíveis entre si.

Neste novo contexto, a Web será capaz de representar associações entre “coisas” que, em princípio, poderiam não estar relacionadas. Para isso, computadores necessitam ter acesso a coleções estruturadas de informações (dados e metadados) e de um conjunto de regras de inferência que ajudem no processo de dedução automática.

Estas regras são especificadas através de ontologias que permitem representar explicitamente a semântica dos dados; no item 4.7.2, abordaremos com mais profundidade este tópico.

O primeiro passo para separar as informações na Web será a substituição da linguagem HTML que tem justamente esta característica de juntar dados e apresentação pela linguagem XML, com características de dar significado aos dados, conforme visto no capítulo 3.

As iniciativas em torno da Web Semântica apontam para que o conteúdo disponível na Web seja codificado, de forma que seja possível o processamento automático pelos computadores. Desta forma, as pesquisas realizadas em mecanismos de busca, por mais complexas que sejam, retornariam apenas o resultado esperado, como acontece nas consultas às bases de dados convencionais. Para isso, é necessário padronizar um mecanismo consistente de metadados.

Os documentos são mais fáceis de localizar e gerir se soubermos alguma coisa sobre eles, como o nome do autor, data de publicação, assunto etc. Este tipo de informação, que define "dados sobre dados", é o que consideramos metadados. Ao disponibilizar um arquivo para download, um exemplo de metadados para este arquivo seria: nome do programa, versão, tamanho do arquivo, informações sobre a licença de uso, plataforma etc.

Segundo Grácio (2002, p.114), metadados podem ser definidos como: “Conjunto de elementos que descrevem as informações contidas em um recurso, com o objetivo de possibilitar sua busca e recuperação”.

Metadados então, são utilizados para descrever as características de recursos e seus relacionamentos. Tradicionalmente, o uso de metadados é associado a sistemas gerenciadores de banco de dados. Na última década, metadados ganharam uma nova dimensão e sua

importância é essencial no gerenciamento e manutenção de data warehouses, mecanismos de busca, ferramentas de software etc.

Na Web, o imenso conteúdo disponível e a heterogeneidade dos recursos evidenciam cada vez mais a necessidade de adoção de padrões para metadados, a fim de aprimorar e facilitar a recuperação da informação. Conforme visto no capítulo 3, documentos HTML ainda representam a maior parte dos objetos acessados via WWW. Geralmente, estes não provêm nenhuma informação extra associada capaz de otimizar o processo de pesquisa; que limita-se a palavras-chaves, indexação de textos ou páginas inteiras, sem acesso a conteúdos específicos. O objetivo da linguagem HTML é simplesmente formatar a exibição de páginas Web.

Descrever o conteúdo e não apenas exibi-lo, é o primeiro passo para a criação da Web Semântica. A seguir, detalharemos algumas tecnologias de base para a criação da Web Semântica, com enfoque nas iniciativas que contribuíram para a padronização de metadados na Web.

4.7 Proposta de estrutura da Web Semântica

Segundo Berners-Lee, Lassila, Hendler (1999), a forma da Web Semântica é representada através da figura a seguir:

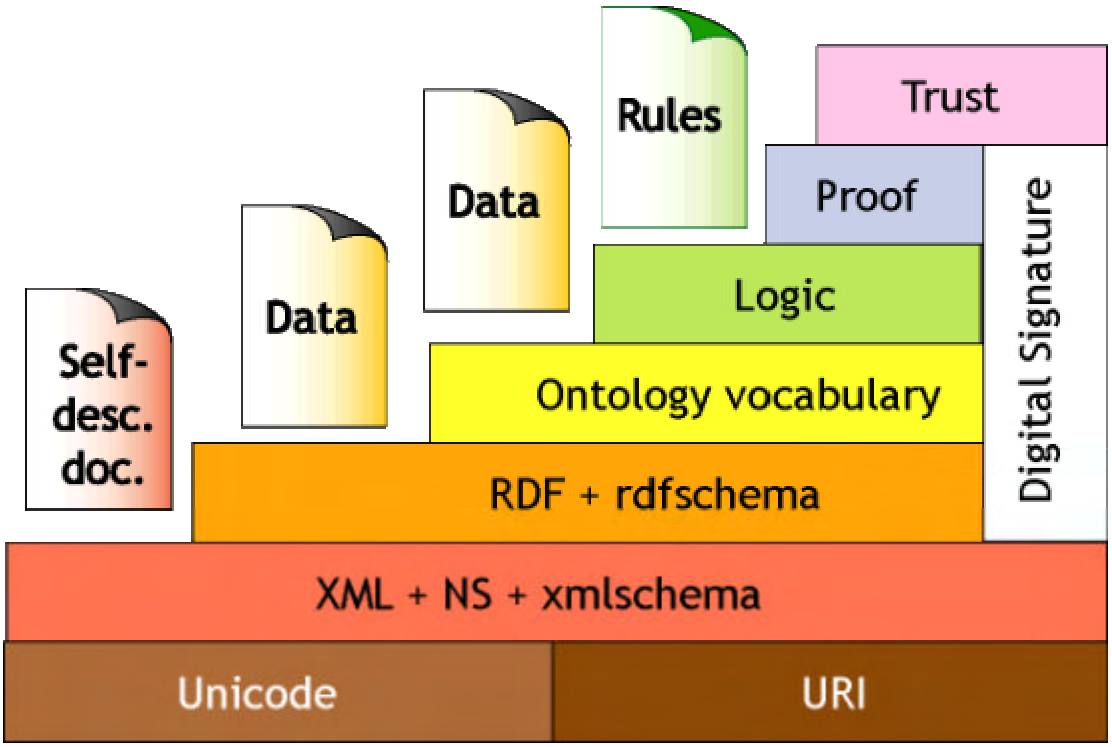

FIGURA 3 – CAMADAS DA WEB SEMÂNTICA

FONTE: HTTP://WWW.W3.ORG/2000/TALKS/1206-XML2K-TBL/SLIDE10-0.HTML. ACESSO EM 20/09/2003.

Na camada base da figura 3, encontramos URI (Uniform Resource Indentifiers) e Unicode que são os padrões para descrição de identificadores universais de recursos e códigos internacionais de dados. Um exemplo de URI é o ISBN (International Standard Book Number).

Para entendermos melhor a segunda camada (XML + NS + xmlschema), utilizamos a seguinte citação de Greenberg (2003),

> XML e mais recentemente schemas de XML facilitam a criação, o uso e a interoperabilidade sintática dos vocabulários de metadados, e o Ns (namespaces), que são identificadores através de URIs, garantem a segurança entre vocabulários de metadados.

A terceira, quarta, quinta e sexta camadas apresentam-se dentro de um grupo chamado assinatura digital (Digital Signature), que é uma

tecnologia que vem se aperfeiçoando cada vez mais, e apresenta-se como um grupo na figura, pois se faz necessário a validação da integridade dos dados que serão utilizados para conclusão das tarefas de agentes dentro da Web semântica.

Na terceira camada, apresentamos o RDF, juntamente com o RDF Schema. A arquitetura RDF foi concebida para descrever metadados sobre recursos.

Segundo Lassila (1999),

> RDF é uma aplicação da linguagem XML que se propõe ser uma base para o processamento de metadados na Web. Sua padronização estabelece um modelo de dados e sintaxe para codificar, representar e transmitir metadados, com o objetivo de torná-los processáveis por máquina, promovendo a integração dos sistemas de informação disponíveis na Web. (tradução nossa)

A especificação de RDF define como descrever recursos em termos de suas propriedades e valores; um processo muito parecido com um Diagrama Entidade Relacionamento.

A quarta camada, Vocabulário Estruturado, pode ser entendida segundo Greenberg(2003), como

> A camada de ontologia [que] representa a veia semântica central de metadados na Web, onde simples descrições para complexos esquemas classificatórios devem ser criados e registados, de modo que os agentes inteligentes possam interpretar dados, fazer inferências e executar tarefas.

A quinta camada apresenta a Lógica que, segundo Greenberg (2003), “[...] é a mesma lógica que utilizamos no desempenho de nossas atividades diárias”. O autor afirma, ainda, que: “Um agente pode derivar uma conclusão lógica (ou a razão) no processo de terminar uma tarefa baseada no que são essencialmente apresentados pelo código de metadados semântico”; isso implica afirmar que um determinado agente

toma decisões lógicas baseado nas descrições encontradas nas estruturas mais inferiores.

A sexta e sétima camadas (Proof e Trust) apresentam-se fora da estrutura lógica e tem por finalidade testar a confiabilidade da informação a ser recuperada, com base, principalmente, na criação de metadados existentes nas camadas inferiores.

A sustentação da Web Semântica é baseada na camada de estrutura com Dados, Metadados, linguagem XML (eXtensible Markup Language) e arquitetura RDF (Resource Description Framework), camada de esquema que aborda a questão de Ontologias e a camada lógica com as Regras de Inferência. A figura 4 a seguir apresenta uma forma mais simples de arquitetura das camadas da Web Semântica (tripé principal):

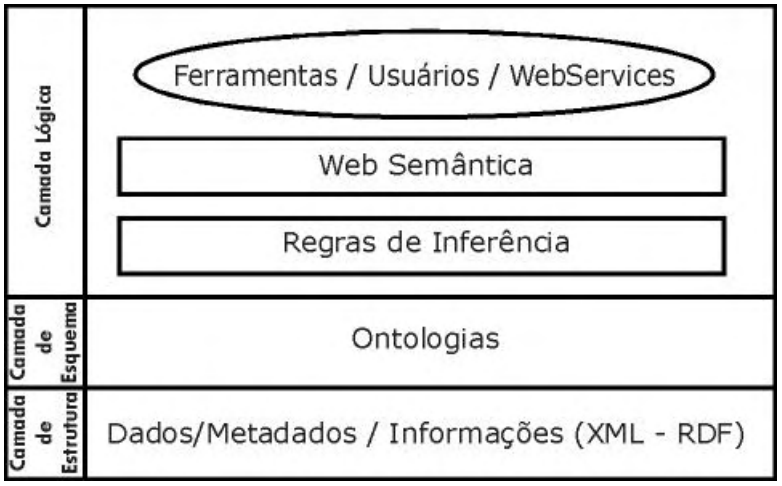

FIGURA 4 – ARQUITETURA MAIS SIMPLES DE CAMADAS DA WEB SEMÂNTICA

Com base nesta figura, abordamos detalhadamente, nas seções seguintes, as camadas de estrutura, de esquema e lógica.

4.7.1 CAMADA DE ESTRUTURA E RDF (RESOURCE DESCRIPTION FRAMEWORK)

Considerando que foram abordados anteriormente os itens metadados e XML, apresentamos a seguir as linguagens RDF (Resource

Description Framework) e RDF Schema (Resource Description Framework Schema).

Segundo Silva e Lima (2002):

> O fundamento da RDF estabelece um modelo básico para descrição de dados que consiste em três tipos de objetos:
>
> - Recursos: um recurso é qualquer coisa descrita em expressões da RDF. Pode ser uma página da Web, um website inteiro ou parte deste. Pode ser também um objeto não acessível via Web, como um livro, uma revista ou um CD. Recursos são sempre especificados por URI's.
> - Propriedades: uma propriedade é uma característica, um atributo ou uma relação utilizada para descrever o recurso.
> - Declarações: uma declaração é um recurso específico com uma propriedade definida mais o valor desta propriedade. Podemos dizer que uma declaração é um recurso mais as propriedades desse recurso e mais o valor dessas propriedades.

Essas três partes individuais são chamadas respectivamente de sujeito, predicado e objeto. Em outras palavras, o modelo básico primitivo da RDF consiste em conjuntos (registros) de objeto, propriedade e valor.

Declaração= Recurso (sujeito) + Propriedades do recurso (predicado) + Valor da propriedade (objeto)

EXEMPLO 34 – DECLARAÇÃO RDF

Considere a seguinte sentença: José Eduardo é aluno do Programa de Pós Graduação em Ciência da Informação, onde:

- "Programa de Pós-Graduação em Ciência da Informação" é o sujeito (recurso);
- "aluno" é o predicado (propriedade);
- "José Eduardo" é o objeto (literal - valor da propriedade).

Esta sentença pode ser representada pelo diagrama abaixo:

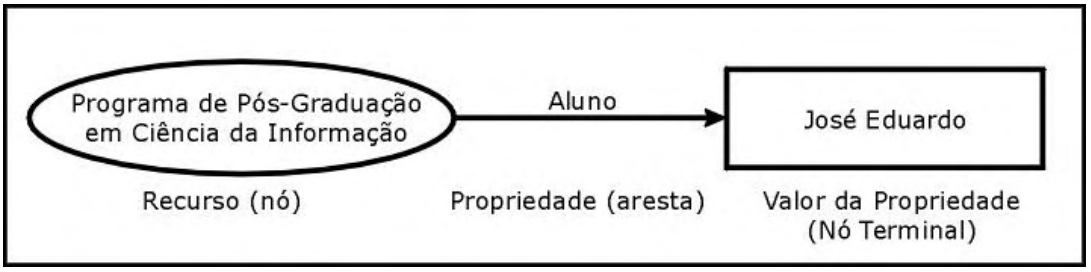

FIGURA 5 – DIAGRAMA 1 - RDF

A orientação da aresta é significante: o arco sempre começa no sujeito (recurso) e aponta para o objeto da declaração (valor da propriedade). O diagrama também pode ser entedido como: O Programa de Pós-Graduação em Ciência da Informação tem como aluno José Eduardo, ou, de uma maneira geral, "<sujeito> TEM <predicado> <objeto>".

Para acrescentar mais características ao objeto da declaração ("José Eduardo"), como por exemplo, seu e-mail pessoal, teríamos a seguinte sentença:

O indivíduo cujo nome é José Eduardo, email <eduardo@unespmarilia.br>, é aluno do Programa de Pós-Graduação em Ciência da Informação.

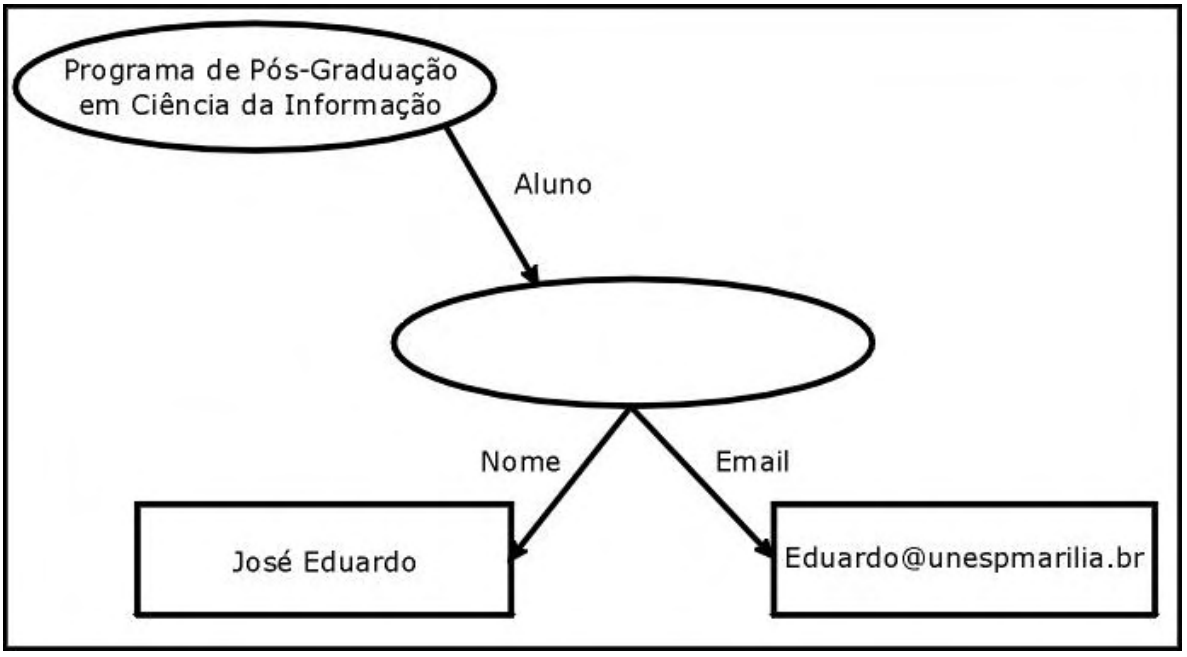

FIGURA 6 – DIAGRAMA 2 - RDF

A diferença entre a sentença (1) e a sentença (2) é que na primeira temos uma declaração com uma única propriedade relacionada a um objeto, um literal. Enquanto que na sentença (2), há uma propriedade estruturada (aluno) que possui duas outras propriedades (Nome e e-mail).

Em RDF, propriedades estruturadas são representadas como outro recurso e como a sentença em questão não dá nome para este recurso, ele é anônimo, e por isso representado por uma elipse vazia. Este diagrama também pode ser lido como: Programa de Pós-Graduação em Ciência da Informação tem alguém como aluno, esse alguém se chama José Eduardo e possui o e-mail eduardo@unespmarilia.br.

O modelo de dados RDF apresentado anteriormente fornece uma abstração para definição dos metadados utilizados. Para efetivamente criar os metadados, é necessária uma sintaxe concreta. A especificação da RDF utiliza a linguagem XML para definir sua sintaxe e o conceito de namespaces para associar cada propriedade com o esquema que a define.

Lassila (1999) relata que a especificação do W3C apresenta duas sintaxes de XML para codificação de um modelo de instância de dados em

RDF: a sintaxe de serialização e a sintaxe abreviada. A diferença mais marcante entre as duas está em como a estrutura do modelo RDF é apresentada. A primeira nos oferece uma estrutura mais completa enquanto a segunda nos oferece uma forma mais compacta.

Para ilustrar, demonstramos a seguir a sintaxe de serialização. Sendo assim, o modelo discutido seria codificado da seguinte forma:

```
01 <?xml version='1.0' encoding='ISO-8859-1'?>
02
03  <rdf:RDF
04   xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
05   xmlns:rdfs = "http://www.w3.org/2000/01/rdf-schema#"
06   xmlns:turmas = "http://www.unespmarilia.br/ci/turmas#"
07  >
08
09  <rdf:Description rdf:about="Programa de Pós-Graduação em CI">
10
11 <turmas:Aluno rdf:ID="http://www.unespmarilia.br/posci/eduardo">
12  <turmas:nome>José Eduardo</turmas:nome>
13     <turmas:email>eduardo@unespmarilia.br</turmas:email>
14   </turmas:Aluno>
15
16  </rdf:Description>
17
18 </rdf:RDF>
```

EXEMPLO 35 – SINTAXE SERIALIZAÇÃO RDF

O documento inicia com a declaração XML na linha 1, uma vez que a RDF é uma aplicação da XML. Esta declaração especifica a versão da XML à qual o documento obedece - neste caso, a versão 1.0, que atualmente é a única versão. O valor "ISO-8859-1" do atributo encoding especifica que caracteres acentuados e o cedilha poderão fazer parte do conjunto de dados de caracteres do documento.

A linha 3 declara que o bloco <rdf:RDF> ... </rdf:RDF> (linha 18) é uma expressão RDF e usa o vocabulário definido pelos espaços de nome declarados nas linhas 4, 5 e 6. Os prefixos de espaço de nome "rdf", "rdfs" e "turmas" referem-se respectivamente à sintaxe padrão da RDF e RDF Schema e a um vocabulário criado especificamente para fornecer informações sobre os alunos de uma turma.

Na linha 9, a tag “description” é usada para indicar qual recurso da Web terá seus metadados descritos. Em nosso exemplo, o recurso descrito é o Programa de Pós-Graduação em Ciência da Informação. Observem que esta tag é finalizada na linha 16 com a marcação </rdf:Description> .

As linhas 11-14 utilizam o prefixo de espaço de nome "turmas" para descrever as informaçoes sobre aluno do programa que está sendo descrito. Esse aluno é identificado por sua página pessoal disponível em: http://www.unespmarilia.br/posci/eduardo. O atributo ID da sintaxe padrão RDF define esta identificação.

As marcações das linhas 12 e 13 definem, respectivamente, o nome e o e-mail do aluno. Os elementos utilizados para aplicar descrições sobre o aluno precisam estar em conformidade com o vocabulário definido em http://www.unespmarilia.br/ci/turmas#.

A proposta RDF é a de flexibilidade para formalização de vocabulários que possam ser processados por máquinas e que ainda sejam legíveis aos humanos, tornando possível especificar semântica para dados baseados em XML de uma maneira padronizável e interoperável.

A estrutura RDF define um mecanismo para descrever recursos que não faça nenhum pré-julgamento sobre um domínio de aplicação particular, nem defina, a priori, a semântica de um domínio de aplicação, descrevendo um mecanismo que seja independente de domínio, mas que permita descrever informações sobre qualquer domínio.

Segundo Silva e Lima (2002),

> A RDF pode ser utilizada em várias área de aplicações da Web: na busca de recursos para melhorar os mecanismos de sites de busca já existentes, em bibliotecas virtuais descrevendo o conteúdo disponível, no comércio eletrônico, principalmente na segurança, em websites particulares, etc. Também é util em outras aplicações que estão fora do escopo da Web, como recursos multimídias em geral, bibliotecas digitais e outras. A RDF em si é uma linguagem simples capaz de fazer relacionamentos entre informações, mas, além disso, é necessário um meio para definição de dados. A RDF Schema foi criada pelo W3C com essa finalidade.

Os esquemas RDF definem o significado, as características e os relacionamentos do conjunto de propriedades dos recursos. Definem também os tipos de recursos que estão sendo descritos. Podem ser entendidos como uma espécie de dicionário onde são especificados os termos que serão utilizados em declarações RDF. O objetivo é estabelecer regras para garantir que os dados estejam sempre em conformidade com elas. A figura 7 a seguir ilustra o contexto de um esquema RDF.

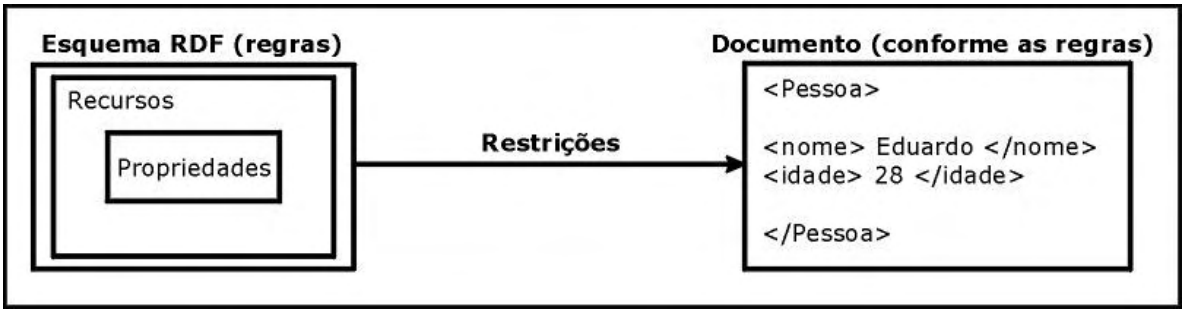

FIGURA 7 – ESQUEMA RDF

Segundo Brickley e Guha (2003), o vocabulário da RDF Schema é definido em um espaço de nome denominado "rdfs" e identificado pela URI.

Conforme Beckett (2003), a especificação também utiliza o prefixo "rdf" para referir-se ao espaço de nome do modelo e sintaxe RDF, identificados pela URI.

Os recursos podem ser instâncias de uma ou mais classes e isto é expresso em um esquema RDF através da propriedade rdf:type. Assim como na programação orientada a objetos, as classes podem ser organizadas de forma hierárquica e os conceitos de herança são aplicáveis aos recursos. Para explicar a relação entre as classes, a especificação da RDF Schema descreve a propriedade rdfs:subclassOf.

A figura 8 mostra o conceito de classe, subclasse e recurso. As classes (e subclasses) são descritas por um retângulo de cantos arredondados e os recursos por pontos negros. As setas partem de um recurso até a classe que o define.

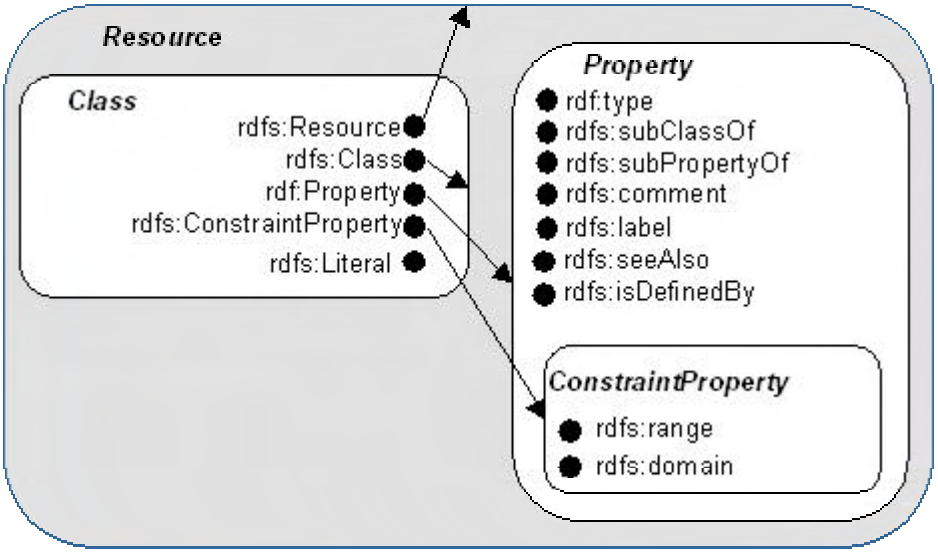

FIGURA 8 – CLASSES RDF

FONTE: BRICKLEY E GUHA(2000)

Convenções:

- Nomes de Classe têm a primeira letra maiúscula (Ex: rdfs:Resource);

- Palavras adicionais ao nome de classe são também maiúsculas (Ex: rdfs:ConstraintProperty);
- Nomes de propriedades não têm a primeira letra maiúscula (Ex: rdfs:domain);
- Palavras adicionais ao nome são maiúsculas (Ex: rdfs:subClassOf).

A figura a seguir (BRICKLEY e GUHA, 2000) ilustra a hierarquia de classes para o esquema RDF, utilizando a representção gráfica de um grafo, onde as classes são representadas por nós.

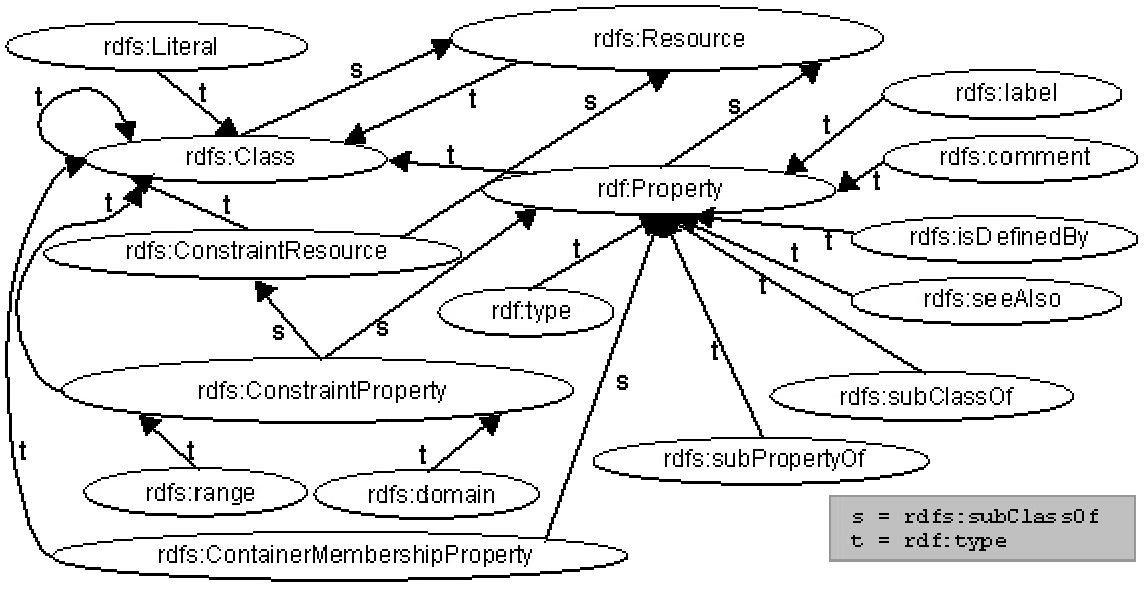

FIGURA 9 – HIERAQUIA DE CLASSES RDF

FONTE: BRICKLEY E GUHA (2000)

Se uma classe é subclasse de outra, então existe um arco rdfs:subClassOf dela para sua respectiva superclasse. Do mesmo modo,

se um recurso é uma instância de uma classe, então existe um arco rdf:type do recurso para a sua respectiva classe.

As classes essenciais de um esquema RDF são:

- rdfs:Resource - representa o conjunto de recursos do modelo básico da RDF. É a mais geral das classes, pois tudo pode ser definido como recurso. Todas as demais classes são subclasses dela;
- rdf:Property - representa o subconjunto de recursos que são propriedades da RDF;
- rdfs:Class - similar à noção de uma classe em linguagens de orientação a objetos, tendo em vista que uma classe RDF pode representar qualquer recurso, tais como websites, pessoas, tipo de documentos ou conceitos abstratos. Quando um esquema RDF define uma nova classe, o recurso que a representa deve ter a propriedade rdf:type, cujo valor é o recurso que define a classe que o contém.

As propriedades descritas a seguir fornecem um mecanismo para expressar relacionamentos entre classes e suas instâncias ou superclasses. São instâncias da classe rdf:Property:

- rdfs:type - indica que um recurso é membro de uma classe e tem todas as características inerentes a um membro dessa classe. Quando um recurso tem uma propriedade rdf:type cujo valor é alguma classe específica, esse recurso é uma instância dessa classe. Um recurso pode ser uma instância de uma ou mais classes;
- rdfs:subClassOf - mostra que uma classe pertence a outra mais abrangente na hierarquia de classes. Uma subclasse pode pertencer a mais de uma classe;

- rdfs:subPropertyOf - é uma instância de rdf:Property e indica que uma propriedade é uma especialização de outra;
- rdfs:seeAlso - especifica que um recurso contém informações adicionais sobre o recurso em questão;
- rdfs:isDefinedBy - é uma subpropriedade de rdfs:seeAlso e especifica o recurso que define o recurso em questão.

O exemplo 36 ilustra a descrição de classes em um esquema RDF que descreve veículos, utilizando a especificação da RDF Schema apresentada em Brickley e Guha (2003).

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<rdf:RDF
  xmlns:rdf = "http://www.w3.org/TR/rdf-syntax-grammar#"
  xmlns:rdfs = "http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#">

  <rdf:Description ID="Veiculo">
   <rdf:type resource="http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#Class"/>
   <rdfs:subClassOf
    rdf:resource="http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#Resource"/>
  </rdf:Description>

  <rdf:Description ID="VeiculoDePassageiro">
   <rdf:type resource="http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#Class"/>
   <rdfs:subClassOf rdf:resource="#Veiculo"/>
  </rdf:Description>

  <rdf:Description ID="Onibus">
   <rdf:type resource="http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#Class"/>
```

```
        <rdfs:subClassOf rdf:resource="#VeiculoDePassageiro"/>
       </rdf:Description>

       <rdf:Description ID="Van">
        <rdf:type        resource="http://www.w3.org/TR/2003/WD-rdf-schema-
20030905#Class"/>
        <rdfs:subClassOf rdf:resource="#VeiculoDePassageiro"/>
       </rdf:Description>

      </rdf:RDF>
```

EXEMPLO 36 – DESCRIÇÃO DAS CLASSES RDF

4.7.2 CAMADAS ESQUEMA

A partir das definições técnicas apresentadas anteriormente, abordaremos nesta Seção a parte mais lógica da informação, dando início ao acesso as ontologias.

Os sistemas atuais para representação do conhecimento são centralizados, exigindo que todos compartilhem, as mesmas definições de conceitos, cada qual tendo um conjunto diferente de regras para fazer inferências sobre os seus dados.

Mas o que são ontologias? Segundo Guizzardi (2000),

> Ontologias fornecem o conhecimento estruturado e uma infra-estrutura para integrar bases de conhecimentos, independentes da implementação e constituem uma ferramenta poderosa para suportar a especificação e a implementação de sistemas computacionais de qualquer complexidade.
>
> Em alguns casos, esse termo é usado apenas como um nome mais rebuscado, denotando o resultado de atividades familiares como modelagem de domínio e análise conceitual. No entanto, em muitos outros casos, as ditas ontologias apresentam algumas peculiaridades como a forte ênfase na

> necessidade de uma abordagem altamente formal e interdisciplinar, na qual a filosofia e a lingüística desempenham um papel fundamental.

Muitas são as definições sobre Ontologias: acreditamos que a que mais se aplica à Web Semântica é a de Jacob (2003) que diz que:

> Ontologias são categorias de coisas que existem ou podem existir em um determinado domínio particular, produzindo um catálogo onde existem as relações entre os tipos e até os subtipos do domínio, provendo um entendimento comum e compartilhado do conhecimento de um domínio que pode ser comunicado entre pessoas e programas de aplicação.

Com a utilização das regras de inferência, as aplicações mais avançadas usam ontologias para relacionar a informação em uma página às estruturas de conhecimento associadas.

Conforme Novello (2002):

> O uso de Ontologias torna possível definir uma infra-estrutura para integrar sistemas inteligentes no nível do conhecimento. O nível do conhecimento é independente do nível de implementação. Ontologias apresentam grandes vantagens como:
>
> • Colaboração: possibilitam o compartilhamento do conhecimento entre os membros interdisciplinares de uma equipe.
>
> • Interoperação: facilitam a integração da informação, especialmente em aplicações distribuídas;
>
> • Informação: podem ser usadas como fonte de consulta e de referência do domínio;
>
> • Modelagem: as ontologias são representadas por blocos estruturados que podem ser reusáveis na modelagem de sistemas no nível de conhecimento.

Novello (2002) afirma, ainda, que

> [...] as ontologias podem servir como uma ferramenta navegacional de consulta para o usuário, fornecendo informação semântica sobre restrições, conceitos e relacionamentos do domínio, mantendo o conhecimento do

> domínio compartilhado entre todos os membros de uma equipe e até mesmo entre equipes geograficamente separadas.

Assim as ontologias permitem a criação de novos conceitos de busca: as ferramentas de consulta poderão buscar e trocar informações precisas dentro de um determinado domínio e até entre domínios distintos, permitindo a criação das regras de inferência que veremos na camada lógica.

4.7.3 CAMADA LÓGICA

Na camada lógica da Web Semântica encontramos as regras de inferência e as aplicações e serviços para Web. Segundo Nunes (2002),

> As regras de inferência são mecanismos que possibilitam inferir, a partir de asserções válidas, expressões válidas, isto é, consistentes com todas as interpretações possíveis.
>
> Dessa forma, é possível concluir-se sobre validade, ainda que não se considere interpretações particulares.
>
> Ou seja, através da sintaxe, conclui-se sobre propriedades semânticas (validade).

Bräscher (2002) relata que

> A Web semântica utiliza-se ainda das ontologias para possibilitar a recuperação de conceitos. Uma ontologia na Web Semântica possui uma taxonomia e um conjunto de regras de inferência. A taxonomia define as classes de objetos e as relações que se estabelecem entre eles. Forma-se assim uma estrutura onde propriedades são atribuídas a determinadas classes e os objetos que lhe pertencem herdam suas características. A solução de ambigüidades e a obtenção de maior precisão na recuperação de informações disponíveis na Web constituem-se numa das principais preocupações dos estudos da Web Semântica. [...] Assim, os motores de busca poderão encontrar páginas que se refiram

> a conceitos específicos e não à todas as páginas nas quais a palavra ambígua é utilizada.

Este processo de aplicação de regras de inferência, para busca de resultados e respostas adequadas na Web, é realizado por softwares agentes, disponibilizados através de serviços de aplicação para Web, possibilitando o aparecimento de novos serviços e tecnologias de informática aplicadas aos processos de localização, descrição e disseminação da informação da Web.

4.8 NOVAS TECNOLOGIAS APLICADAS AOS PROCESSOS DE LOCALIZAÇÃO, DESCRIÇÃO E RECUPERAÇÃO DA INFORMAÇÃO DA WEB

Anteriormente verificamos que muitas são as tecnologias que vêm sendo desenvolvidas para serem aplicadas à Web. Algumas já podem ser implementadas, enquanto outras são apenas conceitos que deverão fazer parte do dia a dia dos usuários em um futuro próximo.

A implementação destas novas tecnologias juntamente com outras que já existiam estão criando um espaço e fazendo surgir novas ferramentas e *web services*.

Apesar das novas tecnologias estarem sendo muito importantes, algumas como a tag <META> já existiam e não eram utilizadas. Assim estaremos, baseados nas tecnologias apresentadas nos capítulos anteriores e em novas tecnologias, analisando o que tem sido feito e o que deve ser feito, para que as comunidades que utilizam a Internet tornem disponíveis trabalhos realmente bem documentados que possam gerar resultados significativos.

4.8.1 WEB SERVICES

Os *Web services* são serviços oferecidos através da Internet e estão crescendo de uma maneira avassaladora, estando diretamente ligados às tecnologias que estão surgindo.

As grandes empresas estão apostando muito na Internet e vêm dividindo com a comunidade acadêmica as pesquisas das novas tecnologias para solucionar problemas que ainda dificultam as transações via Web.

De acordo com Greco (2002, p.29),

> As previsões apontam para uma rápida expansão no uso de serviços Web. A empresa IDC prevê que essa tecnologia vai gerar nos Estados Unidos, 1,6 bilhões de dólares em negócios em 2004, mas esse número deve aumentar para 34 bilhões de dólares em 2007, um crescimento espantoso em apenas 3 anos.

Quando falamos de *Web services* estamos falando de novas ferramentas que estarão disponíveis para serem utilizadas por usuários nos mais variados setores e áreas afins.

Entre os serviços que estaremos estudando neste capítulo estão os mecanismos de busca “inteligentes” que deverão ser implementados nas ferramentas de busca, permitindo que as mesmas produzam resultados significativos aos usuários, entendendo a Web Semântica e realizando consultas com resultados baseados em conteúdo, ou seja, em significados semânticos com utilização de softwares agentes de localização inteligentes.

Os programas agentes ou agentes inteligentes que terão como função coletar conteúdos na Web a partir de fontes diversas, processar a informação e permutar os resultados com outros programas, permitindo através da linguagem expressar inferências lógicas resultantes do uso de regras e informação como aquelas especificadas pelas ontologias.

O principio não está, no entendimento, pela máquina, daquilo que está escrito, mas no reconhecimento de provas escritas na linguagem estabelecida pela ontologia; os programas-agentes, pela inferência lógica, retornam respostas ao que foi requerido, ou agente e consumidor podem alcançar entendimento compartilhado, permutando as ontologias que oferecem o vocabulário necessário para a discussão.

A utilização de recursos para localização de informação também nos leva a observar a utilização da linguagem WebSQL (Query Language for Web), que vem sendo estudada desde 1997 e que é baseada na linguagem de consulta SQL, utilizada nos principais bancos de dados existentes no mercado.

A WebSQL é uma ferramenta de consulta a páginas desenvolvidas utilizando linguagem XML que, a partir da Web Semântica, poderá ser um dos recursos mais utilizados na recuperação de informação na Web. A WebSQL tenta fazer com que a Web seja entendida como se fosse um grande banco de dados relacional e utiliza-se dos novos recursos de armazenamento de informações na Web para fazê-lo.

Entre os novos serviços que aparecem baseados na linguagem XML e RDF, para descrição e armazenamento de informações na Web, encontramos o MARC XML (Machine Readable Cataloging XML) e a linguagem OWL (Ontology Web Language).

O MARC XML assim como o formato MARC (Machine Readable Cataloging) foram desenvolvidos pela Biblioteca do Congresso Americano (Library of Congress – LC). O MARC XML é a versão XML do formato MARC, muito utilizado em catalogação na Ciência da Informação, e deve ser um formato para ser utilizado em locais onde estão sendo construídos ambientes baseados em XML, como por exemplo em repositório institucionais de teses e dissertações e em bibliotecas digitais.

A linguagem OWL, segundo BECHHOFER,

> [...] é também uma linguagem de marcação baseada em XML, para publicação e compartilhamento de ontologias na Web, e foi desenvolvida como um vocabulário extensivo à RDF. (tradução nossa)

Com o advento da Web Semântica, a OWL passa a ser fundamental na construção das ontologias.

As novas tecnologias permitem um processo de recuperação mais eficiente e, de certa maneira, que agrade mais aos usuários. Uma das particularidades que já vem sendo implementada e agrada bastante aos usuários é encontrarem sites que visitam freqüentemente, moldando uma área pessoal única destinada a cada visitante, de acordo com as características de navegação do próprio usuário; é o que se chama de criação de perfis de usuários, através de interações do próprio usuário na Web.

Neste processo vai se criando um perfil de acordo com as últimas visitas, buscas, compras, solicitações e outras ações que o usuário realizou em sua visita a uma determinada área da internet.

Outro processo que já pode também ser encontrado é o de filtragem de informações, baseado em metadados chamado de PICS (Platform for Internet Content Selection), que é uma ferramenta para seleção de conteúdo na Internet.

As normas técnicas, chamadas de PICS, foram desenvolvidas de forma a fornecer aos usuários uma estrutura para classificar suas páginas Web, permitindo que esses rótulos sejam lidos automaticamente e protegendo os usuários de material indesejado, principalmente pornografia e violência.

O texto, Platform for Internet Content Selection (2003), editado pela W3C, mostra que a plataforma PICS planeja um padrão que “permite aos pais e professores controlar as informações que as crianças estão recebendo pela Web” (tradução nossa)

A figura 10 demonstra um serviço de PICS implementado em um sistema de Browser.

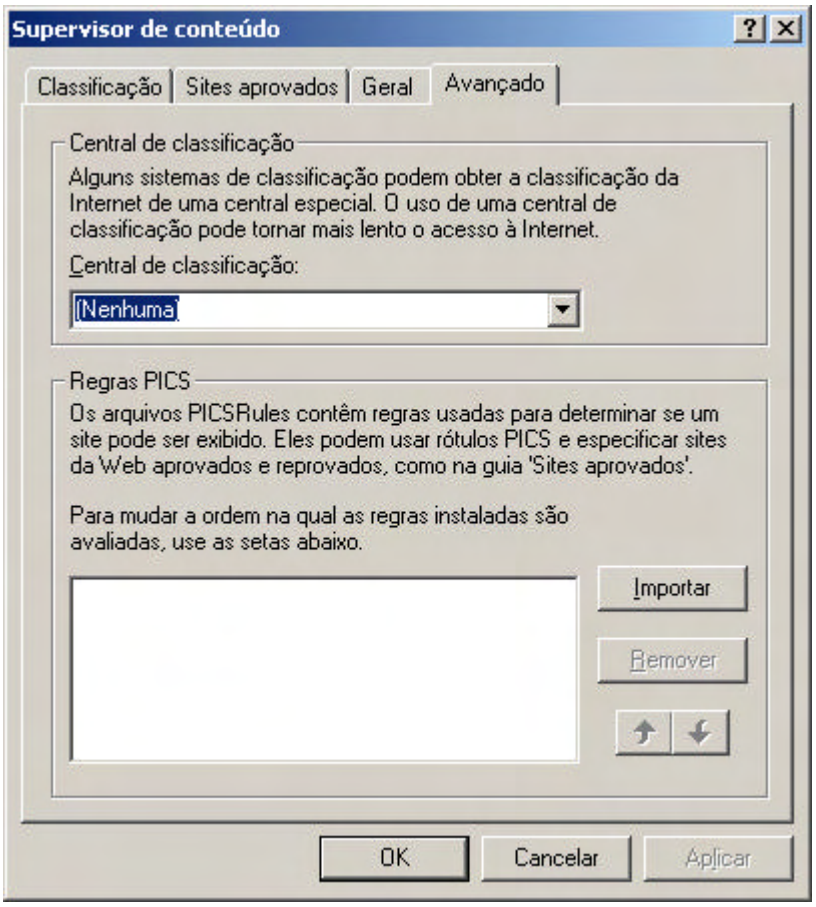

FIGURA 10 – REGRAS PICS

Durante este capítulo apresentamos a Web como repositório de informação e ferramenta de estudo da Ciência da informação, fazendo um histórico sobre a evolução dos serviços e apresentando as novas tendências que vêm aparecendo, possibilitando novos serviços que fazem

ou estarão fazendo parte da vida de todos os usuários da Web num futuro próximo.

CAPÍTULO 5
CONSIDERAÇÕES FINAIS: UM OLHAR PARA O FUTURO

5 CONSIDERAÇÕES FINAIS: UM OLHAR PARA O FUTURO

Estamos vivendo um mundo globalizado, repleto de informações e a Internet vem revolucionando o conceito de proximidade e de tempo, aproximando pessoas e facilitando serviços. Neste contexto, percebemos uma aproximação muito forte das principais Ciências abordadas neste documento, a Ciência da Informação e a Ciência da Computação, principalmente quanto aos objetivos gerais.

Verificamos que a informação na Web é um dos objetos de estudo da Ciência da Informação e que a Internet tem muito a oferecer a esta ciência no tocante à pesquisa e à realização de novos serviços.

Conforme afirma Valentim (2002, p.119),

> [...] o tripé informação, tecnologias da informação e telecomunicações muda a sociedade e, conseqüentemente, muda suas demandas. Nesse sentido, o profissional da informação deve ter uma postura investigativa e crítica, de modo que possa assumir essas mudanças sociais de forma natural. Por fim, a globalização, fenômeno mundial que afeta profundamente as relações sociais e de trabalho, cria novas situações para os profissionais que atuam com dados, informação e conhecimento.

Observamos que o crescimento/desenvolvimento tecnológico está presente nas mais diversas áreas do conhecimento e, cada vez mais, a interdisciplinaridade aparece quando fazemos um estudo, criando uma grande sinergia entre as diversas comunidades de pesquisa; verificamos, assim, que a Ciência da Informação é uma das ciências que mais agrega conhecimento de outros campos de estudo, o que já era dito por Le Coadic (1996, p.22),

> A interdisciplinaridade traduz-se por uma colaboração entre diversas disciplinas, que leva a interações, isto é, uma certa reciprocidade, de forma que haja, em suma, enriquecimento

> mútuo. A forma mais simples de ligação é o isomorfismo, a analogia.
>
> A Ciência da Informação é uma dessas novas interdiciplinas, um desses novos campos de conhecimentos onde colaboram entre si, principalmente, a psicologia, a linguística, a sociologia, a informática, a matemática, a lógica, a estatística, a eletrônica, a economia, o direito, a filosofia, a política e as telecomunicações.

Durante este trabalho verificamos que o tratamento da informação digital está sofrendo mudanças, acarretadas principalmente pelo crescimento da quantidade de informação disponível na Internet e que ferramentas e conceitos estão sendo desenvolvidos, procurando facilitar o trabalho de armazenamento, descrição, indexação e recuperação de informações na Web.

Observamos que as ferramentas de busca são serviços importantes, mas ainda muito limitados, pois descrevem e indexam uma quantidade de informação muito pequena em relação ao tamanho da Web; como produzem resultados, na maioria das vezes, indesejados ou insignificantes, grande parte deste problema é gerado pela forma de armazenamento das informações que se apresentam sem um padrão ou formato que permita a estes serviços um melhor resultado para o usuário.

Para Bueno e Vidotti (2000, p.8)

> [...] recuperar informações na WWW sem uma estratégia e um instrumento adequado significa obter milhares de documentos irrelevantes. Portanto, é imprescindível conhecer os recursos disponíveis pela própria WWW para se ter a resposta desejada.

Silva (2003, p.87) destaca que,

> [...] apesar da evolução tecnológica, ocorrida na última década, as ferramentas de busca ainda encontram alguns problemas ao descrever, indexar e classificar as informações digitais que são inseridas nos seus bancos de dados.

Avaliando a maneira com que as informações são descritas na Web percebemos que é necessária uma padronização e que só, assim, poderemos ter uma recuperação mais objetiva das informações desejadas.

Vidotti (2001, p.7) afirma que

> A Internet possui documentos que se interligam conforme a organização prevista pelos seus criadores. A navegação nesse sistema hipermídia mundialmente construído por milhões de indivíduos pode conduzir o usuário a uma busca incessante de informações, sem garantia de atingir o objetivo principal da busca ou de sua utilização.

Verificamos que a linguagem XML é ponto de partida para padronização de representação de conteúdos na comunicação entre agentes computacionais; é baseado neste conceito que vm sendo discutidos e implementados novos serviços. Observamos, também, que o conceito da linguagem XML produz um efeito especial na publicação de informação, não apenas pela nova forma de produzir informação, mas pela maneira que se apresenta, como uma ferramenta pautada em estudos e técnicas metodológicas que visam a facilitar a troca e recuperação da informação por comunidades distintas.

Segundo Siqueira (2003, p.71),

> Pode-se afirmar que a XML é uma linguagem para criar padrões de comunicação entre sistemas de computadores, o que permitirá a integração tanto da base de dados como de arquiteturas, hardwares e métodos de programação usados, favorecendo a interoperabilidade. Entretanto, a XML sozinha não é nada, ou seja, um arquivo de computador com informações estruturadas segundo a metodologia XML só terá seu valor prático se outras tecnologias estiverem sendo usadas em conjunto.

Com o objetivo de produzir resultados coerentes em um processo de buscas na Web, o caminho a ser seguido é o de fornecer

semântica às informações dispostas na rede; para isso, é necessária a utilização de ferramentas como RDF e acima de tudo ontologias que darão o verdadeiro significado semântico ao conteúdo.

Durante o estudo verificamos o delineamento de uma nova estrutura de informações na Web, chamada de Web Semântica; a publicação de informações, de acordo com esta nova proposta, é uma questão de tempo e, que em breve, essa nova extensão da Web passará a ser um espaço consistente e qualificado de informações dentro da Internet, possibilitando a comunidades acadêmicas a construção de conhecimento a partir de dados confiáveis encontrados na rede.

Entre os principais resultados obtidos com esta pesquisa destacamos a relação, muito próxima, existente entre a Ciência da Informação e a Ciência da Computação. Como graduado na área de computação pudemos perceber principalmente que os estudos e pesquisas da Ciência da Informação estão muito próximos dos vivenciados na graduação em Ciência da Computação.

Pudemos verificar que a Ciência da Informação se pauta na socialização da informação, onde quer que ela esteja, que se importa muito com a criação de padrões para socializar esta informação, carência encontrada na Ciência da Computação que prefere se ater à criação de mecanismos técnicos de processamento de informações, aprofundando muito mais a parte estrutural do que social do projeto a ser desenvolvido.

Segundo Pinheiro (1999, p.177),

> A Ciência da Informação, a Comunicação e a Ciência da Computação formam um triângulo disciplinar altamente dependente da nova ordem tecno-cultura[...] o que poderá, no futuro, levar a formação de uma disciplina com características transdisciplinares, to dipo Infocomunicação.

Percebemos, também, a importância da aproximação entre os profissionais da Ciência da Informação e da Ciência da Computação, que

troquem experiências e idéias, além, claro, das particularidades estudadas em cada uma das áreas; quem sabe possam criar ambientes que permitam a acessibilidade das informações na Web, construindo caminhos e ferramentas que permitam interoperabilidade de informação através de técnicas metodológicas, visando a construção de novos conhecimentos.

Dessa maneira, este trabalho procurou criar um estudo pautado pela linha de Informação e Tecnologia da Ciência da Informação, que possa contribuir para que novos projetos direcionados à informação na Web possam utilizá-lo em busca da interoperabilidade das informações disponíveis na Internet.

REFERÊNCIAS

REFERÊNCIAS

ALMEIDA, M. B. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. *Ciência da Informação*, Brasília, v. 31, n. 2, p. 5-13, maio/ago. 2002.

ANÁLISE comparativa de máquinas de busca para a web brasileira. Disponível em: <http://www.akwan.com.br/mix_analise.shtml>. Acesso em: 20 jul. 2003.

ANDERSON, R. et al. *Professional XML*. Rio de Janeiro: Ciência Moderna, 2001. 1266 p.

BARRETO, A. A. A questão da informação. *São Paulo em perspectiva*, São Paulo, v. 8, n. 4, p. 3-8, out./dez. 1994.

_____. Mudança estrutural no fluxo do conhecimento: a comunicação eletrônica. *Ciência da Informção*. Inf., v. 27, n. 2, p. 168-75, maio/ago. 1998.

BARITE, M. Organización del conocimiento: un nuevo marco teórico-conceptual en bibliotecología y documentación. In: CARRARA, K. (Org.) *Educação, universidade e pesquisa*. São Paulo: FAPESP, 2001. p.35-46.

BAX, M. P. Introdução as linguagen de marca. *Ciência da Informação*, Brasilia, v.30, n.1, p.32-38, jan./abr. 2001.

BECHHOFER, S. et al. *OWL web ontology language reference.* 2003. Disponível em <http://www.w3.org/TR/owl-ref/>. Acesso em: 10 jul. 2003.

BECKETT, D.(eds.) RDF/XML Syntax Specification. 2003.Disponível em: <http://www.w3.org/TR/rdf-syntax-grammar>. Acesso em: 19 set. 2003

BERGMAN, M. K. The deep web: surface hidden value. *Journal of Electronic Publishing*, v.7, n.1, Aug. 2001. Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html#to7b.> Acesso em: 02 out. 2003.

BERNERS-LEE T.; LASSILA, O.; HENDLER, J. The semantic web. *Scientific American*, v. 5, 2001. Disponível em: <http://www.sciam.com/print_version.cfm? articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 19 mar. 2003.

BORKO. H. Information Science: what is? American Documentation, v. 19, n. 1, p. 3-5, Jan. 1968.

BRÄSCHER, M. A ambigüidade na recuperação da informação. *DataGramaZero: revista de ciência da informação*, Rio de Janeiro, v.3, n.1, 2002. Disponível em: < http://www.dgzero.org/fev02/Art_05.htm>. Acesso em: 10 ago. 2003.

BRASIL. Ministério da Ciência e Tecnologia. *Sociedade da Informação*: ciência e tecnologia para a construção da sociedade da informação no Brasil: bases para o Brasil na sociedade da informação: conceitos, fundamentos e universo político da indústria e serviços de conteúdo. São Paulo: Instituto UNIEMP, 1998.

BRICKLEY, D.; GUHA, R. V. (Eds.) *RDF Vocabulary Description Language: RDF schema*. Disponível em: <http://www.w3.org/TR/2003/WD-rdf-schema-20030905/>. Acesso em: 09 set. 2003.

_____. *Resource Description Framework (RDF) Schema Specification 1.0.* 2000. Disponível em: http://www.w3.org/TR/2000/CR-rdf-schema-20000327/. Acesso em: 19 mar. 2003.

BRYAN, M. *Guidelines for using XML for Electronic Data Interchange*, 1998. Disponível em: <http://www.geocities.com/WallStreet/Floor/5815/guide.htm>.

BUENO, M. C.; VIDOTTI, S. A. B. G. Ferramentas de Busca na Internet: para quê, por quê e como utilizá-las? In: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS e DOCUMENTALISTAS - *Informação*: o desafio do futuro, 2., 2000. (1 CD-ROM).

_____. Uso estratégico das ferramentas de busca na Internet. In: SIMPÓSIO INTERNACIONAL "PROF. DR. PAULO TARCÍSIO MAYRINK, 3, 1999, Marília. *Anais...* Marília: Faculdade de Filosofia e Ciências, 1999. p.39-49.

BUSCANDO termos perto de outros. Disponível em: <http://www.cnpq.br/plweb/info/ajuda/nearops.html#xproxop>. Acesso em: 02 out. 2003.

CAMPOS, J.; SANTACHÈ, A.; TEIXEIRA, C. *Visualização de modelos tridimensionais de sistemas de informações geográficas distribuídos baseados na WEB*. In: BRAZILIAN WORKSHOP ON GEOINFORMATICS, 1., 1999, Campinas. Anais... Campinas, 1999.

CENDÓN, B. V. Ferramentas de busca na web. *Ciência da Informação*, Brasilia, v.30, n.1, p.39-49, jan./abr. 2001.

DEROSE, S. et al. (Ed.) *XML Linking Language (XLink) Version 1.0*, 2001. Disponível em: <http://www.w3.org/TR/xlink/>. Acesso em: 15 maio 2003.

EUCLIDES, M. L. *A indústria da informação no Brasil: o acesso à informação*. 10f. 2000. Trabalho apresentado à Disciplina Acesso à Informação do curso de Especialização das Tecnologias em Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília.

FERNANDES, G. C. O objeto de estudo da Ciência da Informação. *INFORMARE*: Cadernos do programa de pós-graduação em Ciência da Informação, Rio de Janeiro, v. 1, n. 1, p. 25-30, jan./jun. 1995.

GODOY, J. Entendendo um documento SGML. In:_____. *Introdução à linguagem SGML*. 2000. Disponível em: <http://lie-br.conectiva.com.br/godoy/sgml-2.html>.

GRÁCIO, J. C. A. *Metadados para descrição de recursos da Internet:* o padrão Dublin Core, aplicações e a questão da interoperabilidade. 127 f. 2002. Dissertação (Mestrado em Ciência da Informação). Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2002.

GRECO, M. *De carona com web services*. Info Corporate, São Paulo, n.1, p.27-29, dez. 2002.

GREENBERG, J. The Semantic Web: more than a vision. *Bulletin for the American Society for Information Science and Technology*. V. 29, n.4, p.6-7, apr./may, 2003.

GUIMARÃES, C. *Introdução a Linguagens de marcação HTML, XHTML, SGML, XML*. Campinas: Unicamp, 2002. Disponível em: <http://www.dcc.unicamp.br/~celio/inf533/docs/markup.html>.

GUIMARÃES, J. A. C. *Perspectivas de ensino e pesquina em organização do conhecimento em cursos de Biblioteconomia do Mercosul: uma reflexão*. (Trabalho apresentado no V Encuentro de Investigadores de Bibliotecologia y Ciencia de la Información de Iberoamerica y el Caribe (EDIBCIC), Granada, 21-25 feb. 2000.

GUIZZARDI, *G. Uma abordagem metodológica de desenvolvimento para e com reuso, baseada em ontologias formais de domínio*, 2000. Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, Vitória, 2000.

HOLZNER, S. *Desvendando o XML*. Rio de Janeiro: Campus, 2001. 858p.

IBGE. Pesquisa nacional por amostras de domicílio: PNAD, 2002. Disponível em: < http://www.ibge.gov.br/home/presidencia/noticias/10102003pnad2002.shtm>.

IBICT. *Projeto CCN*: módulo 3, tutorial. 1997. Disponível em:<http://www.intelecto.net/cn-ead/mod3.htm>. Acesso em: 06 junho 2003.

JACOB, E. K. Ontologies and the semantic web. *Bulletin for the American Society for Information Science and Technology*. V. 29, n.4, p.19-22, abr./maio 2003.

KADE, A. M. *Linguagens de consulta para documentos XML*. Porto Alegre: UFRGS, 1999. 36 p.

Le COADIC, Y. *A ciência da informação*. Brasília: Briquet de Lemos, 1996. 119 p.

LÉVY, P. *Cibercultura*. Tradução C. I. da Costa. São Paulo: Ed. 34, 1999. 264p.

LIMA, L. O. *Mosaic revolucionou o acesso à rede*. Grupo Estado, 2000. Disponível em: <http://www.estado.estadao.com.br/edicao/especial/internet/interne2.html>.

LYMAN, P.; VARIAN, H. R. *How much information?* 2000. Disponível em: <http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html>.

MACHLUP, F.; MANSFIELD, U. ed. *The study of information*: interdisciplinary messages. New York, John Wiley & Sons, 1983. 743p.

MARCONDES, C. H.; SAYÃO, L. F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. *Ciência da Informação*, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em: <http://www.ibict.br/cionline/300301/3030401.htm>. Acesso em 2 maio 2003.

_____. Documentos digitais e novas formas de cooperação entre sistemas da informação em C&T. Ciência da Informação, Brasília, v. 31, p. 42-45, set./dez. 2002.

MEGGINSON, D. (Ed.) *XML Information Set Requirements*, 1999. Disponível em:<http://www.w3.org/TR/NOTE-xml-infoset-req>.

MESSAROS, J.D. *Folhas de estilo CSS nível 1*. 2000. Disponível em: <http://geocities.yahoo.com.br/doc2web/W3C/CSS1/REC-CSS1-19990111.html#ref2>.

NOVELLO, T. C. *Ontologias: sistemas baseados em conhecimento e modelos de banco de dados*. Disponível em: < http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/artigo_taisa .pdf>. Acesso em 15 maio 2003.

NUNES, M. G. V. *Regras de inferência*. Disponível em: <http://www.icmc.sc.usp.br/~gracan/download/sce5832/lpo4.html>. Acesso em: 10 jul. 2003.

PINHEIRO, L. V. R. Campo interdisciplinar da Ciência da Informação: fronteiras remotas e recentes. In:_____. (Org.) *Ciência da Informação, Ciências Sociais e interdisciplinaridade*. Brasilia: IBICT, 1999. p.155-182.

PLATFORM for Internet Content Selection (PICS). 2003. Disponível em: <http://www.w3.org/PICS/>. Acesso em: 08 agosto 2003.

PORQUE usar o Google. Disponível em: <http://www.google.com.br/intl/pt-BR/why_use.html>. Acesso em: 10 set. 2003.

RAMALHO, J. A. *XML*: teoria e prática. São Paulo: Berkeley, 2002. 146 p.

RAMALHO, J. C. L. *Anotação estrutural de documentos e sua semântica*. Disponível em: http://www.di.uminho.pt/~jcr/XML/publicacoes/teses/phd-jcr/src/book1.htm. Acesso em: 15 set. 2003.

RAY, E. T. *Aprendendo XML*. Rio de Janeiro: Câmpus, 2001. 372 p.

ROWLEY, J. *A Biblioteca eletrônica*. Tradução de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos / Livros, 2002.

SANTAREM SEGUNDO, J. E.; VIDOTTI, S. A. B. G. Organização da informação na Web: a busca na qualidade do armazenamento e da recuperação com a utilização de XML e RDF. In: V SIMPÓSIO EM FILOSOFIA E CIÊNCIA, 5, 2003, Marília. *Trabalho e conhecimento:* desafios e responsabilidades da ciência: anais eletrônicos. Unesp Marília Publicações, 2003. CD-ROM.

SANTAREM SEGUNDO, J. E.; VIDOTTI, S. A. B. G.; FUSCO, E.; BREGA, J. R. F. Linguagem XML como base na busca da interoperabilidade e organização da informação. In: V SIMPÓSIO EM FILOSOFIA E CIÊNCIA, 5, 2003, Marília. *Trabalho e conhecimento:* desafios e responsabilidades da ciência: anais eletrônicos. Unesp Marília Publicações, 2003. CD-ROM.

SANTOS, P. L. V. A. C. As novas tecnologias na formação do profissional da informação. In: VALENTIM, M. L. *Formação do profissional da informação*. São Paulo: Polis, 2002. p. 103- 116.

SEMANTIC web activity statement. Disponível em: <http://www.w3.org/2001/sw/ Activity>. Acesso em: 15. fev. 2003.

SILVA, F. R. da. *Ferramentas de Busca na Internet*: um estudo do Google, Yahoo! e Metapesquisador. 2003. 103f. Monografia (Trabalho de Conclusão de curso) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – Unesp. Marília. 2003.

SILVA, G. C.; LIMA, T. S. RDF e RDFS na infra-estrutura de suporte à web semântica. *Revista Eletrônica de Iniciação Científica*. SBC, v.2, n.2, mar. 2002. Disponível em: <http://www.sbc.org.br/reic/edicoes/2002e1/cientificos/Edicao_Marco_2002_Artigo09_Resumo.htm>. Acesso em: 15 fev. 2003.

SIQUEIRA, M. A. *XML na Ciência da Informação*: uma análise do MARC 21. 109 f. 2003. Dissertação (Mestrado em Ciência da Informação). Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2003.

SMITH, J. W.; BARRETO, A. A. Ciência da informação: base conceitual para a formação do profissional. In: VALENTIM, M. L. *Formação do profissional da informação*. São Paulo: Polis, 2002. p. 9-23.

TAKAHASHI, T. (org.) *Sociedade da informação no Brasil*: livro verde. Brasília: Ministério da Ciência e Tecnologia, 2000. 203p.

TAYLOR, R. Professional aspects of information science and technology. *Annual Review of Information Science and Technology*, v. 1, p. 15-40, 1966.

VIDOTTI, S. A. B. G.. *O Ambiente hipermídia no processo de ensino-aprendizagem*, 2001. 125f. Tese (Doutorado em Educação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP, Marília, 2001.

WEBOPEDIA. *Search engine*. 2003. Disponível em: <http://www.webopedia.com/TERM/s/search_engine.html>. Acesso em: 18 de jul. 2003.

W3C CONSORTIUM. *Extensible Markup Language (XML)*, 2003. Disponível em: <http://www.w3.org/XML>. Acesso em: 02 maio 2003.

W3C DOM WORKING GROUP. *W3C Document Object Model*. 2002. Disponível em: <http://www.w3.org/DOM>. Acesso em: 05 maio 2003.

WURMAN, R. S. Ansiedade de informação: como transformar informação em compreensão. 5.ed. São Paulo: Cultura, 1995. 380p.

YAHOO! Brasil: perguntas mais frequentes. Disponível em : <http://br.yahoo.com/info/faq.html>. Acesso em: 10 set. 2003.

APÊNDICE

APÊNDICE

Marcação	Função	Características
<HTML> </HTML>	Delimita o documento (início e fim)	Esta marcação deve ser inserida imediatamente no início e no final de todo arquivo HTML
<HEAD> </HEAD>	Cabeçalho	O cabeçalho de um arquivo html é uma área para inserção de informações que não serão visíveis na página HTML, tais como “title” e comentários
<TITLE> </TITLE>	Título do documento	Nome a ser atribuído ao documento HTML, mas que não será visualizado na página. Deve sempre aparecer na área definida como “head” do documento
<! >	Comentário	Indicação de quaisquer comentários, tais como nome do autor, data de criação, software utilizado para autoria, etc. Não é visualizado na página HTML (só quando solicitado o “source file”) Também deve aparecer na área definida como: “head” do documento.
<BODY> </BODY>	Corpo do documento	Delimita o corpo do documento. Vem imediatamente abaixo da área definida como “head”. Esta marcação deve ser “fechada” somente ao final do documento, imediatamente seguida da marcação </html>
<BODY	Imagem de fundo	Insere uma imagem como fundo da

Marcação	Função	Características
BACKGROUND="*"> </BODY>	* = arquivo imagem ou url de arquivo imagem	página (mais ou menos como uma "marca d'água). O arquivo deve estar em formato .gif . Este arquivo pode estar na mesma máquina (anotar sua path/nome), ou em outra máquina (indicar por URL)
<BODY BGCOLOR="#rrggbb"> </BODY>	Cor padrão de fundo #rgb = código de cores rgb (ver código de cores RGB neste documento)	Define uma cor de fundo padrão para a página. Esta cor pode variar de monitor para monitor.
<BODY TEXT="#rgb" LINK="#rgb" VLINK="#rgb" ALINK="#rgb"> </BODY>	Cores do texto do documento #rgb = código de cores rgb (ver código de cores RGB neste documento)	Define a cor para: text = texto normal da página link = links da página vlink = links consultados na página alink = links ativados na página
<Hy> </Hy>	Títulos ou cabeçalhos na página y = número de 1 a 6	Existem seis níveis de cabeçalho na página (não confudir com cabeçalho do documento = head): , , etc. Sendo H1 o maior e H6 o menor nível. Atenção, marcações Hy definem tamanho relativo do texto, e os colocam em destaque (normalmente negrito)
<P>	Parágrafo	Insere uma linha em branco entre

Marcação	Função	Características
		dois parágrafos
 	Quebra de linha	Faz uma quebra de linha.
<A HREF="**"> texto	Âncora para hiperlink referencial ** = URL ou nome de arquivo	Define um link. O link vai aparecer em destaque na página (normalmente outra cor e sublinhado)
<A HREF="#**"> texto	Âncora interna ou para seção específica em outro documento #** = código ou palavra chave	Abre uma âncora para um outro trecho dentro de uma mesma página, ou para um trecho específico em outro documento. 1. Trecho na mesma página Após a âncora de hiperlink referencial, deve aparecer, entre aspas, o caracter # e uma palavra ou código chave. Este mesmo código ou palavra deverá aparecer no "ponto de chegada" deste link interno, como uma âncora de nome (veja próximo item). 2. Link para trecho específico em outro documento Deve ser efeito exatamente da mesma forma, sendo que , antes do caracter #, deverá aparecer o nome do arquivo que se pretende 'ligar', ou sua url completa.
<A NAME="** ">	Âncora de nome	Este é o ponto de chegada de uma âncora interna a um documento ou para trecho específico em outro

Marcação	Função	Características
	**= código ou palavra chave	documento. O código ou palavra chave deve ser idêntico àquele do ponto de partida. Não altera a visualização do texto.
<A HREF=“MAILTO: alguem@algumlugar.br ”>	Envio de email para endereço especificado	Permite que se crie um link que ao ser selecionado abrirá uma tela de composição de mensagem eletrônica a ser enviada para o endereço digitado após MAILTO:
<UL> </UL>	Delimita lista não numerada	Deve ser escrita ao início e ao final da lista. Cada item da lista é antecedido da marcação <LI>
<OL> </OL>	Delimita lista numerada	Deve ser escrita ao início e ao final da lista. Cada item da lista é antecedido da marcação <LI>
<LI>	Item da lista	Serve para identificar cada item de uma lista numerada ou não
<DL></DL>	Delimita lista de definições	Deve ser escrita ao início e ao final da lista. Cada item da lista é antecedido da marcação <DT> ou <DD>, conforme explicado a seguir.
<DT>	Entrada de título em lista de definições	Insere um título em uma lista de definição. A entrada de título vai aparecer alinhada à esquerda da página.

Marcação	Função	Características
		Normalmente vem seguido de um item do tipo <DD>
<DD>	Entrada de definição em lista de definições	Insere uma definição (antecedido ou não por <DT>). Este item vai aparecer numa margem mais interna à página (como se tivesse uma tabulação antes). Fora de listas de definição pode ser utilizado para produzir este mesmo efeito (Uma “tabulação)
<PRE></PRE>	Texto pré-formatado	Mantém a exata formatação do texto digitado.
<B></B>	Negrito	Insere negrito no texto
<I></I>	Itálico	Insere itálico no texto
<ADRESS> </ADRESS>	Endereço	Quanto à visualização, normalmente, apenas coloca o trecho em itálico.
<IMG SRC=“**”>	Inserção de imagem ** = nome ou URL de arquivo imagem.	Marcação para inserir uma imagem na página. Esta imagem deve estar preferencialmente em formato .gif. Pode estar na mesma máquina (apontar com o nome ou path completa) ou em outra máquina (apontar com URL)
<IMG ALING=“” SRC=“**”>	Alinhamento de imagem ** = nome ou URL	Opcional. Define o alinhamento de uma imagem na página. Aceita os seguintes valores:

Marcação	Função	Características
	de arquivo imagem.	TOP = alinha o texto com o alto da figura MIDDLE = alinha o texto com o meio da figura BOTTOM = alinha o texto com o rodapé da figura RIGHT = alinha a figura à direita da tela LEFT = alinha a figura à esquerda da tela. Faz ainda com que o texto que esteja ao lado contorne a figura.
<HR>	linha horizontal	Insere linha horizontal no texto
<HR SIZE=n>	Largura da linha horizontal N= número	Opcional Extensão opcional que define a largura da linha
<HR WIDTH=n%>	Ocupação da tela da linha horizontal N= número	Opcional Define o quanto da tela uma linha vai ocupar.

Além das tags acima descritas podemos encontrar todas as outras que fazem parte da versão 4.01 da linguagem HTML, especificada pelo W3C Consortium, acessando o site: http://www.w3.org/TR/html401/