



**UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
DEPARTAMENTO DE BIBLIOTECONOMIA
CURSO DE BIBLIOTECONOMIA**



Jacqueline de Araújo Cunha

WEB SEMÂNTICA: “o estado da arte”

Orientadora: Prof^a Msc. Rildecil Medeiros

NATAL-RN

2006.2

JACQUELINE DE ARAÚJO CUNHA

WEB SEMÂNTICA: “o estado da arte”

Monografia apresentada à disciplina Monografia, ministrada pela professora Maria do Socorro de Azevedo Borba para fins de avaliação da disciplina e como requisito parcial para a conclusão do curso de Biblioteconomia do Centro de Ciências Sociais Aplicadas da Universidade Federal do Rio Grande do Norte.

Orientadora: Prof^ª Msc. Rildecil Medeiros

NATAL-RN

2006.2

Divisão de Serviços Técnicos

Catálogo da Publicação na Fonte. UFRN / Biblioteca Central Zila Mamede

C972w Cunha, Jacqueline de Araújo.

Web Semântica : o estado da arte / Jacqueline de Araújo Cunha. - Natal, RN, 2006.

74f.

Orientadora: Rildecil Medeiros.

Monografia (Graduação em Biblioteconomia). -
Universidade Federal do Rio Grande do Norte. Centro de
Ciências Sociais Aplicadas.

1. Linguagem documentária – Internet - Monografia. 2.
Web Semântica – Monografia. 3. Metadado – Monografia.
4. Ontologia – Monografia. I. Medeiros, Rildecil. II. Título.

RN/UF/DEBIB.

CDU 025.4:004.738.5

JACQUELINE DE ARAÚJO CUNHA

WEB SEMÂNTICA: “o estado da arte”

MONOGRAFIA APROVADA EM ____/____/2006

BANCA EXAMINADORA

Prof^a. MsC Rildecil Medeiros
(Orientadora)

Prof^a. MsC Maria do Socorro de Azevedo Borba
(Prof^a da Disciplina)

Prof^a MsC Mônica Marques de Carvalho
(Membro da Banca)

Ao meu saudoso pai (*in memoriam*),
minha maravilhosa mãe e minhas amadas
irmãs...

Agradeço

À minha família, em especial a minha Mãe e irmãs Janaina e Edna, pelo apoio incondicional, sem vocês nada seria possível,

Ao amigo Humberto Gonçalves, pela disposição com que me ajudou durante atividades acadêmicas,

Aos colegas de sala e em especial as companheiras de trabalho Mônica Karina, Gerlany Rodrigues, Juliana Fonteles e Ogna Pereira, por todas as festas, alegrias, tristezas, conversas e principalmente aprendizado compartilhados. Guardo vocês no coração. Vocês são especiais!

Ao amigo e companheiro de grandes empreitadas, Gustavo César, por tudo que me ensinou e também por toda confiança em mim depositada. Tu és um gigante!

Aos amigos conquistados nos encontros estudantis, Julio Rei (PB), Marcia Rodrigues (PE), Gerlandy Leão (MA), Roosewelt Lins (MA), Rodrigo Meireles (BA), por todas as risadas, conversas, e em especial aos dois últimos pelas ricas contribuições a este trabalho,

Aos amigos Helena e Renildo, pelo apoio que me foi dado no início do curso. Obrigada de coração!

À UFRN, ao DEBIB, a BCZM, e seus colaboradores, pelo apoio concedido e por ter me proporcionado a participação em atividades de grande relevância para minha formação,

Aos que fazem o CPD da BCZM, Lúcia Maranhão, Wellington Silva, Bruno e Tarcísio, por todo apoio e presteza com que sempre me atenderam

A minha orientadora, Rildecil Medeiros, por ter aceitado o convite para orientar este trabalho e por ter me dado todo suporte necessário para seguir até o fim, não somente no que diz respeito à construção deste trabalho, mas principalmente pelo emocional, sempre com uma mensagem de otimismo e confiança, és maravilhosa!

A todas as professoras do DEBIB, pelos ensinamentos e em especial as professoras Mônica Carvalho, e Socorro Borba pelas orientações,

À bibliotecária Elizabeth Sachi Kanzaki Ribeiro, que para mim foi uma verdadeira mestra, e toda a equipe da Biblioteca da Escola de Música da UFRN, por terem me recebido com tanto carinho e também pelos ensinamentos durante o estágio supervisionado,

E finalmente a Deus, a quem agradeço a saúde, a inteligência e a oportunidade de ter vivido até este momento, conhecido estas pessoas, e aprendido com todas elas!

“Uma longa jornada começa com um simples passo”.

Provérbio chinês

RESUMO

Aborda a *Web Semântica* no contexto do ciberespaço. Objetiva principalmente caracterizar a *Web Semântica* a partir do estado da arte, com vistas aos problemas inerentes ao processo de recuperação da informação na *Web*. Utiliza como metodologia a pesquisa bibliográfica bem como processo de análise e discussão textual. Observa que essa área do conhecimento encontra-se em plena evolução e, conseqüentemente, um campo promissor de investigação. Constata que apesar dessa evolução, a literatura nacional ainda é incipiente embora, em nível internacional, caracterize-se como uma área do saber com uma vasta literatura. Conclui que tornou-se possível perceber as dificuldades atinentes à representação e recuperação da informação está atrelado à polissemia lexical e a variação de sentido em diferentes línguas no contexto da *Web*. Recomenda novos estudos nessa área do saber que privilegie a linguagem, o conceito, a ontologia e o metadado.

Palavras-chave: Ciberespaço. *Web Semântica*. Linguagem de marcação. Ontologia. Metadado.

ABSTRACT

It approaches the Semantic Web in the cyberspace context. Its goals to characterize the Semantic Web by “the state of the art”, with sights to the inherent problems to the process of information retrieval in the Web. It uses as methodology the bibliographical research as well as analysis process and discuss of texts. It observes that this field of knowledge meets in full evolution and so a promising inquiry area. It evidences that although this evolution national literature still is incipient even so, at an international level, is characterized as an area of knowing with a vast literature. It concludes became possible to perceive related difficulties to the representation and information retrieval is linked to the lexical polissemic and the variation meaning of different languages on web context. It recommends new studies in this field of knowledge that privileges the language, the concept, the ontology and the metadata.

Keywords: Cyberspace. Semantic Web. Markup language. Ontology. Metadata.

LISTA DE ABREVIATURAS E SIGLAS

AACR – Anglo-American Cataloguing Rules

ARPA - Advanced Research Projects Agency

CERN - Conseil Européen pour la Recherche Nucléaire

CG – Comitê Gestor da Internet

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

DAML - Darpa Agent Markup Language

DARPA - Department of Advanced Research Projects Agency

DC - Dublin Core

DCMI - Dublin Core Metadata Initiative

DTD – Data Type Document

EMBRATEL – Empresa Brasileira de Telecomunicações

FAPESP – Fundação de Pesquisa do Estado de São Paulo

HTML - Hyper text Markup Language

HTTP - Hyper Text Transfer Protocol

IEMML – Information Economy Meta Language

IP - Internet Protocol

IPTO - Information Processing Techniques Office

IPTO - Information Processing Techniques Office

LD – Linguagem Documentária

LNCC – Laboratório Nacional de Computação Científica

MARC – Machine Readable Cataloguing

MCT – Ministério da Ciência e Tecnologia

NCSA – National Center for Supercomputer Activit

NSF – National Science Foundation

OIL - Ontology Inference Layer

OML - Ontology Markup Language

OWL - Ontological Web Language

RDF - Resource Rescription Framework

RNP – Rede nacional de Pesquisa

SGML - Standard Generalize Markup Language

TCP/IP – Transmission Control Protocol/ Internet Protocol

TIC – Tecnologia de Informação e Comunicação

URI – Universal Resource Identifier

URL – Universal Resource Locator

W3C - World Wide Web Consortium

WWW - World Wide Web

XHTML – Extensible Hyper text Markup Language

XML - Extensible Markup Language

XOL - *Ontology eXange Language*

SUMÁRIO

| | | |
|----------|---|-----------|
| 1 | INTRODUÇÃO..... | 11 |
| 2 | O CIBERESPAÇO: contextualizando o objeto de pesquisa | 17 |
| 2.1 | A INTERNET E A <i>WEB</i> | 26 |
| 2.1.1 | Abordagem histórica. | 27 |
| 2.1.2 | Internet no Brasil. | 31 |
| 3 | <i>WORLD WIDE WEB</i> E SEUS PRECURSORES | 35 |
| 3.1 | LINGUAGENS DE MARCAÇÃO. | 39 |
| 3.2 | RECUPERANDO A INFORMAÇÃO NA <i>WEB</i> | 42 |
| 4 | <i>WEB SEMÂNTICA</i>: “o estado da arte”. | 50 |
| 4.1 | A ESTRUTURA DA <i>WEB SEMÂNTICA</i> | 54 |
| 5 | METODOLOGIA. | 63 |
| 6 | CONSIDERAÇÕES FINAIS. | 67 |
| | REFERÊNCIAS. | 69 |

Capítulo 1

Introdução

1 INTRODUÇÃO

A sociedade atual tem apontado cada vez mais para o crescimento exponencial da geração de produtos e serviços de informação. Com isso, a *Web Semântica* tem exercido um papel relevante no estabelecimento de padrões tecnológicos e semânticos para a recuperação da informação, apresentando-se assim de forma transdisciplinar em diversas áreas do conhecimento, ou seja, Ciência da Computação, Ciência da Informação, Engenharia Elétrica, Linguística e Filosofia, sendo as duas primeiras as que estão desenvolvendo trabalhos de forma mais intensa com essa temática no país.

Em busca da compreensão do desenvolvimento da *Web Semântica*, muitos estudiosos têm demonstrado interesse em envidar esforços no processo de geração de novos conhecimentos que contribuam para o seu fortalecimento, uma vez que os impactos na geração e recuperação da informação têm se tornado imensuráveis.

Com isso, esse tema ao ser abordado apontou para a necessidade de uma incursão pela literatura, no sentido de melhor compreender a *Web Semântica*, por meio do seu “estado da arte”. E, dessa forma, elaborar uma monografia como elemento de avaliação para conclusão do curso de Biblioteconomia, na Universidade Federal do Rio Grande do Norte (UFRN).

Assim, a partir da concepção de novos conceitos e de áreas emergentes do saber, apresentados e discutidos, no decorrer do referido curso, em especial na Disciplina de Temática III e no âmbito dessa universidade, foi suscitado o interesse em realizar um estudo que abordasse o processo de indexação. Isto porque, a questão sobre o processo de indexar na Internet, inegavelmente como um grande repositório informacional, ainda apresenta grandes desafios no controle efetivo do que é disponibilizado na rede.

Convém destacar, que a problemática da recuperação da informação e o estabelecimento de padrões tecnológicos e semânticos no contexto do ciberespaço passou a constituir-se a partir do surgimento da *Web*.

Então, diante dessa problemática e da complexidade da polissemia lexical surgiram os seguintes questionamentos para a elaboração deste estudo, ou seja:

- 1) Como identificar a produção científica gerada acerca dessa temática?
- 2) Como caracterizar a *Web Semântica*?
- 3) Qual a importância da *Web Semântica* para os usuários de informações no ambiente digital?

Para responder a estes questionamentos foram estabelecidos alguns objetivos, os quais nortearam o processo de pesquisa e análise das informações. Assim, seu objetivo geral foi caracterizar a *Web Semântica* a partir do seu “estado da arte”, tendo em vista o processo de recuperação da informação no ciberespaço.

Desse modo, teve como objetivos específicos abordar conceitualmente a *Web Semântica*, verificar qual o seu propósito e os possíveis benefícios que poderão ser disponibilizados por essa ferramenta para os usuários da rede mundial de computadores e, por fim, analisar a sua importância com foco interdisciplinar no processo de recuperação da informação.

Dentre as várias metodologias e tecnologias já desenvolvidas com vistas à recuperação da informação na rede, é possível que a *Web Semântica* seja a mais revolucionária. Esta, por sua vez, vem sendo liderada pela *World Wide Web Consortium* (W3C). Tal projeto busca alcançar um elevado nível dos resultados de buscas na rede através do desenvolvimento de padrões tecnológicos e semânticos visando à comunicação entre homem-máquina, de forma que, mesmo sem deter maiores conhecimentos acerca de estratégias de busca, qualquer usuário seja capaz de recuperar a informação de forma mais precisa. Isto porque tal tecnologia pretende permitir que a máquina “compreenda” a necessidade de informação de quem a busca.

Neste sentido, faz-se necessário que os profissionais da informação cada vez mais compreendam essa tecnologia para que possam contribuir de maneira cada vez mais efetiva, uma vez que ao longo dos anos, estes profissionais

já desenvolveram várias metodologias e ferramentas as quais lhes proporcionaram grande *know-how* no tocante ao processo de recuperação da informação.

Desse modo, surgiu o interesse de abordar este tema, que ora configura-se como uma área ainda pouco discutida, em alguns cursos de graduação de biblioteconomia do país. Porém de indiscutível relevância para o profissional bibliotecário, visto que a Internet consolida-se, cada vez mais, como importante fonte de mediação do processo de comunicação informacional e, conseqüentemente, da recuperação da informação nesse contexto. Apesar disso, os desafios são inegáveis para a gestão de conteúdo informacional e a sua veiculação.

Outra razão que justifica sobremaneira o desenvolvimento de um estudo dessa natureza seria a semelhança entre os mecanismos utilizados na *Web Semântica* para a recuperação da informação com os já utilizados em ambientes analógicos. Isto posto, pelo fato também do *boom* informacional exigir novas habilidades e competências no processo de representação e recuperação da informação, independente da forma do documento, do seu registro gráfico e o ambiente da unidade de informação. De modo geral, estes buscam estabelecer um processo de comunicação que ocorra entre o binômio - usuário-máquina -, através de linguagens de representação do conhecimento.

Com isso, pode-se deduzir que no âmbito da Internet usa-se, muitas vezes, um vocabulário livre para representar conteúdos informacionais, nesse mundo oceânico de informações. Nesse sentido, percebe-se a falta de uma abordagem de caráter metodológico e conceitual que privilegie o uso de linguagem documentária, enquanto instrumento de representação da informação. Isso tem acarretado, de modo geral, uma série de dificuldades no resgate de conteúdos por parte do usuário, que na maioria das vezes desconhece estratégias de busca que possibilitem um maior índice de pertinência e relevância dos resultados obtidos nas suas pesquisas.

Cabe ressaltar, que esse tema já vem sendo pesquisado há algum tempo por diversos profissionais que trabalham com a informação, tais como os informáticos, lingüistas e cientistas da informação. E isto se deve ao fato de que, muito embora a Internet torne acessível uma grande quantidade de informações, ela

também contribui para que informações realmente úteis não sejam recuperadas facilmente. Isto se dá em função da grande quantidade de informações que são produzidas e disponibilizadas a todo o momento, o que causa a geração de um verdadeiro "lixo" informacional cibernético.

Notadamente, em pesquisas práticas pode-se perceber que há diversos motores de busca. Estes, por conseguinte, visam promover uma melhor recuperação da informação, embora este processo ainda não tenha atingido um nível de satisfação muito alto com a utilização dessa ferramenta de busca.

Por isso, este estudo é de caráter investigativo e interdisciplinar, uma vez que necessário se torna uma melhor compreensão de abordagens conceituais de outras áreas do saber. Pois, o mesmo se constitui como um elo que busca uma interação com a ciência e a tecnologia, no sentido de prover a documentação, a informação e, ao mesmo tempo, uma maior inserção do bibliotecário neste contexto. E, ainda, este último como um profissional que tem por missão o tratamento da informação para sua disseminação e uso.

Nesse sentido, é *míster* destacar que a existência de uma brecha digital passa a oportunizar estudos que caracterizem melhor o conteúdo disponível na rede. Pois, acredita-se que a polissemia lexical, o padrão do metadado e a falta de um rigor para com o uso de linguagem documentária no processo de representação da informação, o que tem acarretado problemas no seu processo de busca e recuperação na *web*.

Na sociedade atual, a informação ganhou *status* de insumo básico para a construção do conhecimento e, ainda, para tomada de decisão. Com isso, este bem intangível, este bem econômico seja ele na sociedade como um todo, como também nas organizações, exige que todos os profissionais envolvidos no processo de organização e difusão da informação envidem esforços para promover o acesso e uso às fontes informacionais de forma rápida e qualitativa. Pois, dessa forma estarão contribuindo para construção da inteligência coletiva que possibilitará um desenvolvimento mais hegemônico entre os países e as pessoas.

Partindo desta afirmativa, podemos inferir que a rede mundial de computadores configura-se como um importante canal de transmissão e

compartilhamento de informação/conhecimento, com grandes projeções para o futuro.

No sentido de estruturar este trabalho, fez-se necessária a constituição dos capítulos que se seguem.

Este capítulo introdutório apresenta uma abordagem geral dos objetivos e justificativas que levaram à elaboração da pesquisa, bem como aspectos metodológicos que o viabilizaram.

O segundo capítulo, intitulado, o ciberespaço: contextualizando o objeto de pesquisa, consistiu na caracterização da *Web*. Isto permitiu contextualizar o tema através de uma abordagem histórica acerca do surgimento da internet e da *Web* onde foi descrito o processo através do qual foi possível o surgimento do ciberespaço, bem como alguns dos problemas e necessidades que se deram neste contexto, no que diz respeito ao processo de construção do conhecimento.

O terceiro capítulo discorreu acerca dos precursores da *Web* onde se buscou mostrar os seus primeiros idealizadores e também proporcionar uma visão do amadurecimento das idéias que culminou com a sua criação por Tim Berners-Lee nos anos 90.

O quarto capítulo foi dedicado a explicar a estrutura da *Web Semântica* bem como delinear o seu estado da arte, objetivo principal deste trabalho.

O quinto capítulo diz respeito à metodologia adotada que subsidiou o desenvolvimento deste estudo.

O sexto capítulo consiste das considerações finais onde são colocadas algumas reflexões acerca do tema, bem como os resultados que foram alcançados com esta pesquisa.

Portanto, após esta apresentação de cunho introdutório passa-se então a apresentar o referencial teórico que serviu de base a este estudo.

Capítulo 2

Ciberespaço:

contextualizando o objeto de pesquisa

2 O CIBERESPAÇO: contextualizando o objeto de pesquisa

Na atual Sociedade, a informação configura-se como um elemento de fundamental importância, a qual promove o desenvolvimento em todo e qualquer segmento que dela faça uso, como se percebe no dia a dia.

A importância deste elemento nos dias atuais contribuiu para mudança de paradigmas em todo o mundo e ensejou um novo modelo de sociedade, ou seja, a sociedade da Informação. Esta, por sua vez, em conjunto com o desenvolvimento das Tecnologias de Informação e Comunicação (TIC's), em especial com o advento da Internet, faz com que se veja surgir a cibercultura, ao que Pierre Lévy (1999) assim denominou-a, ao fazer uma alusão à cultura que surge no ambiente cibernético, ou ainda, no ciberespaço.

O termo “Ciberespaço” é de origem americana e foi cunhado em 1984 pelo escritor William Gibson em seu romance de ficção científica *Neuromancien*. Nesse livro, o termo Ciberespaço designa o universo das redes digitais. A partir desse romance, o termo foi imediatamente empregado pelos usuários e criadores de redes digitais.

O ciberespaço tem a sua importância pautada no fato de que ele disponibiliza grandes estoques de informação, elemento sobre o qual a sociedade constrói o conhecimento, bem como promove a interatividade. Esta é conceituada por Barreto (1999, p. 4) como sendo uma representação

da possibilidade de acesso em tempo real pelo usuário a diferentes estoques de informação: às múltiplas formas de interação entre usuário e as estruturas de informação contidas nesses estoques. A interatividade modifica a relação do usuário com o tempo da informação. Reposiciona em nova perspectiva os acervos de informação, como memórias auxiliares de plantão, o acesso à informação e a sua distribuição.

Porém, o mesmo autor faz uma ressalva atentando para o fato de que os repositórios informacionais representam estoques potenciais de conhecimento, e estes, por sua vez, só se efetivam a partir de uma "ação de comunicação mutuamente consentida entre a fonte (os estoques) e o receptor." (BARRETO, 1999, p. 4)

Nesse sentido, na ambiência do espaço cibernético, as informações encontram-se dispersas e em volumes imensuráveis constituindo-se num verdadeiro oceano informacional. Tal fenômeno se dá por suas características de liberdade de publicação, autonomia das fontes e controle descentralizado (CAMPOS, 2005). E, ainda, a linguagem natural utilizada neste ambiente, não torna suficiente para representar conteúdo de documentos com facilidade.

Ainda assim, esta linguagem continua sendo a mais democrática no ciberespaço e também bastante utilizada, visto que representa a forma como uma determinada comunidade se comunica e principalmente pelo fato de que ela reflete a cultura de determinadas comunidades.

Porém, segundo Cintra et al (2002) a linguagem natural (LN) não se aplica ao processo de recuperação da informação, no que tange à indexação, porque sua utilização neste processo seguramente levará à incompreensão e à confusão, devido a fenômenos naturais como a redundância, a ambigüidade, a polissemia e as variações idioletais que são próprias dessa linguagem.

Para que o processo de comunicação ocorra num ambiente de informações desterritorializadas, se faz necessária a utilização de uma linguagem formal, explícita e compartilhada, tais como as linguagens documentárias.

Pois, de acordo com Gardin et al (1968) apud Cintra et al (2002, p. 35):

Uma linguagem documentária (LD) é um conjunto de termos, providos ou não de regras sintáticas, utilizadas para representar conteúdos de documentos técnico-científicos com fins de classificação ou busca retrospectiva de informações.

Assim sendo, a solução para a recuperação da informação na rede é apontada na literatura como sendo uma ação articulada que vislumbre a problemática sob dois aspectos, o da representação dos conceitos e outro do suporte tecnológico que viabiliza a indexação e mecanismos de busca e recuperação da informação.

Diante disso emerge a necessidade de um trabalho de cooperação que aponte para uma dimensão interativa do trabalho do bibliotecário, do lingüista e também do Informático como atores coadjuvantes no processo de geração e

mediação no processo de representação, busca e recuperação da informação disponível na *Web*.

E é neste contexto de caos que surge a *Web Semântica*, a qual "representa a evolução da *Web* atual. Esta visa a "fornecer estruturas e dar significado semântico ao conteúdo das páginas da *Web*, criando um ambiente onde agentes de software e usuários possam trabalhar de forma cooperativa." (MARCONDES, 2005, p. 23).

De acordo com Ferneda (2003), para a construção da *Web Semântica*, se faz necessário adotar linguagens que permitam não apenas a definição de dados através de marcações, mas que também possibilitem descrever formalmente estruturas conceituais que possam ser utilizadas pelos agentes de indexação dos mecanismos de busca.

Pois, se partirmos da premissa de que

"[...] a informação cumpre papel decisivo na mudança dos destinos da humanidade, uma vez que ela está diretamente ligada ao conhecimento e ao desenvolvimento de cada uma das áreas do saber, já que todo conhecimento começa por algum tipo de informação [...]" (CINTRA, et. al, 2002, p. 10)

Podemos inferir que a possibilidade de uso da *Web* de forma mais qualitativa viabilizará uma maior velocidade ao processo de construção do conhecimento e este, por sua vez, promoverá uma aceleração no processo de desenvolvimento humano. O que para, Pierre Lèvy (BIREME, 2006), consistirá no próximo passo da humanidade, onde neste futuro ele vislumbra o surgimento de uma inteligência coletiva, viabilizada pelo compartilhamento de saberes proporcionado pelas novas Tecnologias de Informação e Comunicação (TIC's).

Esta comunicação entre um sistema de informação e usuários se dá a partir do estabelecimento de uma metodologia que implicará na adoção de linguagens documentárias para a representação de conteúdos no processo de indexação. Grande desafio se apresenta quando se pretende criar tal mecanismo para promover uma melhor utilização das informações veiculadas na rede.

Na infovia, a recuperação qualitativa da informação dar-se-á mediante uma rede multidimensional de modelos conceituais, a qual parece ser viável através das ontologias. Estas se apresentam como tecnologia chave para a geração da *Web Semântica*, correspondendo a um entendimento comum e compartilhado sobre um domínio capaz de ser comunicado entre homens e aplicações. (CAMPOS, 2005)

Pierre Lèvy afirma que a *Web* possivelmente trará soluções para o problema científico da gestão do conhecimento, o problema tecnológico da indexação pragmática e semântica, o ciberespaço e o problema sócio-político de governança da inteligência coletiva. (BIREME, 2006).

Para melhor visualizar este processo, cita-se a Internet, onde ao passo em que disponibiliza a informação ela também dificulta o seu acesso, devido à forma com que está sendo inserida no meio virtual.

Para que cumpram seu papel transformador, as informações precisam ser acessadas no tempo e espaço no qual se fazem necessárias, para que dessa forma possa promover o seu desenvolvimento, bem como permita a seus usuários uma utilização mais otimizada de todos os recursos possíveis através da *Web*.

Nesse sentido, a rede mundial de computadores, representada nesse contexto pela Internet, apesar de todas as suas facilidades, por si só, não garante a recuperação das informações, se as mesmas não estiverem organizadas e dispostas em mecanismos que permitam a sua localização de forma eficiente e eficaz. Isto se justifica pelo fato da inteligência artificial ainda não substituir o homem, enquanto autor conteudista no universo *on-line*. E, ainda, a linguagem natural não ser adequada o suficiente para representar conteúdo de documentos com facilidade no ciberespaço.

Essa grande massa de informações, sem padrões de estruturas, ou indexação faz com que a busca por algo específico torne-se uma tarefa árdua para os usuários dessas informações, pois os motores de buscas, comumente utilizados, muitas vezes trazem informações que sequer se aproximam do conteúdo que atenda às necessidades de quem as buscam.

[...] o modelo estrutural da *Web* tem apresentado alguns problemas que podem comprometer a sua evolução. O crescimento além das expectativas e sem controle levou a uma situação em que hoje existe uma quantidade elevadíssima de informação disponibilizada nos repositórios existentes, mas praticamente sem nenhuma organização. Percebe-se, daí, que não há uma estratégia abrangente e satisfatória para a indexação dos documentos e sua conseqüente recuperação.

Visando a buscar soluções para esses problemas, surgem organismos internacionais que através de Consórcio e estabelecendo parcerias buscam a elaboração de modelos estruturais para a padronização da Internet. A exemplo temos a W3C Consortium, uma associação internacional formada por organizações membros do consórcio, pessoas e o público em geral, que trabalham conjuntamente para desenvolver padrões *Web*.

Esse esforço tem apoio também de muitos pesquisadores de vários países. Essas ações convergem para a criação da *Web Semântica*, na tentativa de estabelecer uma forma de se navegar na *Web* não apenas por palavras-chave, mas por unidades conceituais que possam nos ajudar a encontrar documentos de uma forma transdisciplinar e contextual.

Dentre os pesquisadores engajados neste processo pode-se ressaltar mais uma vez o filósofo Pierre Lévy, o qual vem realizando pesquisas ao longo dos últimos 15 anos, "elaborando um extenso mapeamento de conceitos na área das ciências humanas, visando o desenvolvimento de ferramentas que possam auxiliar as pessoas na busca por informações no ciberespaço." (BIREME, 2006)

Desse modo, o estudo de Lévy tem o propósito de:

[...] favorecer aquilo que seria uma Consciência Reflexiva Global, o mapeamento de conceitos desenvolvido por Lévy (IEML – Information Economy Meta Language) propõe uma espécie de comunicação translinguística, no sentido de que os vários domínios de conhecimentos das ciências humanas e sociais poderiam estar aí representados. A diferença dessa metalinguagem é que ela foi desenvolvida pensando-se em sua aplicação no domínio da informática. Trata-se, portanto de uma linguagem que, ao mesmo tempo em que pode ser compreendida por cientistas sociais, pode igualmente ser traduzida por computadores. É essa propriedade do IEML que leva o filósofo a acreditar que poderemos superar a diversidade de linguagens formais nas ciências humanas, inventando os meios que podem aproximar contextualmente seus diversos domínios. (BIREME, 2006)

A *Web Semântica*, criada a partir da pretensão de embutir inteligência e contexto aos códigos das linguagens de marcação XML (sXtensible Markup

Language) utilizados na confecção de páginas da Internet, tem a finalidade de melhorar a interação de *softwares* com estas páginas e também possibilitar um melhor uso por parte dos usuários (SOUZA; ALVARENGA , 2004)

Após o disposto percebe-se a importância da *Web Semântica* para a biblioteconomia e os usuários de informações veiculadas pela Net, visto que esta visa potencializar a acessibilidade das informações disponibilizadas, bem como promover o desenvolvimento deste importante canal de comunicação e divulgação da produção literária, seja ela científica ou não.

Atualmente, o ciberespaço configura-se como um novo ambiente social e tecnológico de expressão, informação e transações econômicas. Neste âmbito, interagem pessoas de todos os países, de todas as culturas e linguagens, de todas as idades e profissões fornecendo, buscando e utilizando informações, por meio de uma rede mundial de computadores interconectada pela infra-estrutura de telecomunicações que permite à informação o processamento e transmissão digital.

Este novo ambiente dispõe a informação e o conhecimento em um espaço e estado contínuos de modificação, em função de sua plasticidade e fluidez, permitindo a interatividade e organizando o conhecimento em forma horizontalizada.

Para Pierre Lévy (1999, p.17),

O ciberespaço é o novo meio de comunicação que surge da interconexão mundial de computadores. O termo especifica não apenas a infra-estrutura material da comunicação digital, mas também o universo oceânico de informações que ele abriga, assim como os seres humanos que navegam e alimentam esse universo. Quanto ao neologismo, cibercultura especifica aqui o conjunto de técnicas (materiais e intelectuais), de práticas, de atitudes, de modos de pensamento e de valores que se desenvolvem juntamente com o crescimento do ciberespaço.

A partir da afirmação de Lévy podemos inferir que o ciberespaço constitui-se no novo ambiente que surge com a Internet e a *Web*, sendo a primeira o suporte tecnológico, ao passo que a segunda constitui-se no principal serviço oferecido pela rede, o qual possibilita a interação do homem com este novo meio de comunicação.

Sua importância no contexto atual reside principalmente no fato de que, através da Internet, podemos acessar conteúdos informacionais a qualquer instante, bastando para isso, termos acesso ao aparato tecnológico necessário, nesse caso um computador ligado à rede.

Segundo Pimenta (2002) "a informação é o sangue da vida moderna". Pois, se constitui em insumo e produto de desenvolvimento de países, organizações e pessoas e, portanto é a partir da apropriação das informações que se torna possível a construção do conhecimento.

Partindo da premissa de que a informação é o elemento que promove o desenvolvimento e a Internet, juntamente com a *Web*, um veículo que possibilita o acesso ilimitado à mesma, a qualquer lugar e hora, fez com que internet fosse vislumbrada por alguns teóricos como sendo a panacéia para as mazelas do mundo moderno, pois se tornava possível um desenvolvimento mais democrático e hegemônico.

O fascínio exercido pelas novas tecnologias da informação gerou, em alguns autores, não só a expectativa de que a expansão econômica da segunda metade da década de 1990 seria perene, como também criou em alguns a impressão de que a Internet poderia representar uma forma de romper todas as barreiras de comunicações entre os países e as pessoas. Notadamente, a expansão das redes de comunicação pela Internet, segundo o pensamento dominante, seria responsável pela inclusão de toda a população mundial no universo das informações e, enfim, por uma suposta "Democracia Global", ou seja, pessoas de qualquer parte do globo, bem como os países poderiam então fazer uso do poderoso remédio da ignorância. (MATTOS, 2003)

Assim sendo, o ciberespaço por tratar-se de um espaço constituído de um grande manancial de informações, aponta para a necessidade de estudar esse novo ambiente de forma a utilizar o seu potencial de maneira mais efetiva. Pois, ao mesmo tempo em que a internet torna possível essa interação entre produtores e usuários de informação, o processo de comunicação, neste ambiente apresenta inúmeras barreiras pelo fato de que não existe um controle efetivo do que é disponibilizado na rede, bem como sobre a utilização da linguagem natural na

representação dos conteúdos que dificulta a sua recuperação. Cabe ressaltar, que a todo o momento são inseridas milhares de informações na Internet, cada qual com suas características e linguagens que são próprias, tais como textos, imagens, sons, dentre outros.

Desse modo, pode-se então afirmar que a problemática da recuperação da informação no ciberespaço pode ser analisada sob dois aspectos. Por um lado, seria pela facilidade proporcionada pela tecnologia que permite que pessoas, que não possuem conhecimento específico em programação ou tecnologias mais aprofundadas, consigam publicar materiais informacionais na Internet. Fato este que se explica através de uma leitura sobre o processo de construção da Internet. Por outro lado, o uso de linguagens inadequadas para representação das informações. Isto tem provocado ruído no processo de comunicação entre homem e máquina, o qual será abordado num outro capítulo.

Cabe ressaltar, que várias são as iniciativas de teóricos e profissionais que convergem para o estabelecimento de uma nova ordem na rede. Porém a mais visível e significativa parece ser a da *W3C (World Wide Web Consortium)*, liderada por Tim Berners-Lee e que tem como objetivo principal construir padrões, com a ajuda de profissionais do mundo inteiro e de linguagens que viabilizem um uso mais otimizado e proveitoso do que é produzido na *Web*. Nesse sentido, o projeto mais ambicioso e que envolve a colaboração de pesquisas em diversas universidades do mundo inteiro é a *Web Semântica*, cujo propósito é dar um significado ao que é veiculado na rede, possibilitando assim, uma melhor interação entre a máquina e o usuário.

Segundo Campos (2005), a *Web Semântica* encontra-se ainda em seus primórdios e por isso ainda é considerada com certa descrença por muitos. Para essa autora,

[...] a *web semântica* pretende fazer com que atividades executadas por usuários, tais como: fazer busca por produtos, consultar a conta bancária, consultar horário de voo de uma empresa aérea, etc, possam também ser executadas por programas. Tais softwares ajudariam a automatizar tarefas mais complexas e a decidir, ao longo das atividades intermediárias, qual serviço utilizar, além de obter e fornecer informações, quando necessário, até completar a tarefa especificada. (CAMPOS, 2005, p.56)

Isso sinaliza também para um projeto bastante ousado e revolucionário, uma vez que promete marcar uma nova etapa no processo de comunicação veiculado pela rede. Com isso, percebe-se que o fruto da *Web Semântica* será o estabelecimento de um sistema de recuperação da informação capaz de recuperar a informação não só pela sua forma gráfica, como também pelo seu conteúdo semântico, como sua própria denominação já sugere.

Porém, antes de uma abordagem mais profunda sobre os aspectos teóricos dessa temática, necessário se faz conhecer um pouco como se deu o surgimento da Internet e da *Web*. Isto posto, para entender melhor uma das razões pelas quais se instaurou o caos informacional no ciberespaço e também o contexto para o qual a *Web Semântica* foi idealizada. Para tanto, abordar-se-á a seguir os conceitos de Internet e *Web*, bem como a trajetória que culminou no seu surgimento.

2.1 A INTERNET E A WEB

De acordo com Ferreira (2001 ,p. 397) o termo Internet significa "rede de computadores de âmbito mundial, descentralizada e de acesso público, cujos principais serviços oferecidos seriam o correio eletrônico e a *Web*", que também quer dizer rede na língua portuguesa.

O termo rede associado aos conceitos de Internet e *Web* pressupõem a existência de unidades interligadas, sejam computadores interligados, informações ou textos, aos quais podemos denominar hipertexto e, também, pessoas. Esta rede baseia-se num modelo de comunicação onde as possibilidades de interconectividade parecem infinitas.

A Internet certamente constitui-se no maior e mais poderoso meio de comunicação da atualidade e, ainda, sem precedentes. Pois, através de sua estrutura possibilita a comunicação em tempo real com qualquer parte do globo, desde que exista uma máquina ligada à rede.

Porém, tal comunicação global veio consolidar-se realmente com a *Web*, pelo fato desta ter proporcionado uma maior possibilidade de interação do

homem com a nova mídia. A *Web*, por conseguinte, ainda ajudou na consolidação do mesmo enquanto novo espaço dimensional, onde as possibilidades de utilização se mostram cada vez mais diversificadas, fazendo surgir ao que Castells (1999) denominou de Sociedade em Rede. Esta, por sua vez, se desenvolve de forma exponencial na ambiência do Ciberespaço, através dos avanços no desenvolvimento das Tecnologias de Informação e Comunicação (TIC's), da interconectividade proporcionada por esta tecnologia, pelos produtos e serviços de informação oferecidos e, conseqüentemente, pelo crescente uso que se tem feito deste meio de comunicação. Além disso, é notória a importância da geração de conteúdo, da cooperação e do compartilhamento de informações para o desenvolvimento da *Web*.

Entretanto, para que todo esse processo tenha se iniciado, um primeiro passo foi dado para o seu surgimento. Assim sendo, passa-se então a abordar de forma breve os seus aspectos históricos.

2.1.1 Abordagem histórica

A Internet também conhecida como Rede Mundial de Computadores, surgiu em meados das décadas de 60 e 70, com outra denominação, no contexto da guerra fria entre as duas maiores potências da época, os Estados Unidos e a ex-União Soviética.

A sua primeira denominação foi ARPANET. Mais tarde seria chamada de Internet, pelo fato de ser um projeto menor que surgiu no âmbito de um dos departamentos da Agência de Projetos em Pesquisas Avançadas (ARPA - *Advanced Research Projects Agency*). Esta agência foi criada pelo Departamento de Defesa dos EUA no ano de 1958, para mobilizar recursos provenientes fundamentalmente do mundo universitário, com o fim de alcançar a superioridade tecnológica militar sobre a União Soviética, que acabava de lançar o seu primeiro satélite artificial, Sputnik, em 1957.

O departamento responsável pelo desenvolvimento da ARPANET foi a Divisão de Técnicas de Processamento de Informação (IPTO - *Information Processing Techniques Office*). O objetivo deste departamento, segundo Joseph

Licklider, seu primeiro diretor, era estimular a investigação no campo da informática interativa.

De acordo com Castells (2004, p. 26) "a construção da ARPANET justificou-se como um meio de repartir o tempo de trabalho on-line dos computadores entre os vários centros de informática interativa e grupos de investigação da agência”.

O IPTO para estabelecer uma interatividade na rede de computadores, baseou-se numa tecnologia revolucionária de transmissão de telecomunicações (*packet-switching*)¹, desenvolvida de forma independente por Paul Barran, então funcionário da Rand Corporation, e por Donald Davies, do Laboratório de Física da Grã-Bretanha.

A proposta da Rand Corporation ao Departamento de Defesa, era a construção de um sistema de comunicação militar, baseado no desenho de uma rede de comunicação flexível e descentralizada, capaz de sobreviver a um ataque nuclear.

Na década de 60, em especial, no ano de 1969, com a tecnologia *packet-switching* já incorporada a ARPANET, os primeiros nós da rede encontravam-se no âmbito de diversas universidades americanas, tais como Universidade da Califórnia e Utah.

Uma vez estabelecida a primeira versão da Internet, o próximo passo foi estender a tecnologia para a conexão com outras redes que foram surgindo. Tal possibilidade acabou por introduzir o conceito de rede das redes.

Devido ao sucesso do sistema criado pela ARPANET, as redes passaram a voltar-se para a área de pesquisas científicas das universidades. Com isso, a ARPA começou a ter dificuldades em administrar todo este sistema, devido ao grande e crescente número de localidades universitárias nela contidas.

¹ De acordo com o Michaelis o termo *packet-switching* significa comutação de pacotes, ou seja, método de envio de mensagens ou dados em pacotes de tamanho uniforme, processamento e roteamento de pacotes em vez de fluxos de *bits*.

Foi então que "em 1983, o Departamento de Defesa, preocupado com possíveis violações do seu sistema de segurança, decidiu criar a rede MIL-NET, destinada exclusivamente a funções militares. A ARPANET converteu-se em ARPA-Internet e destinou-se exclusivamente à investigação."(CASTELLS, 2004, p. 28)

Até então, a comunicação entre as máquinas se dava através de IP (*Internet Protocol* - Protocolo da Internet), porém esta era uma linguagem limitada de comunicação entre máquinas. Para que as redes de computadores pudessem comunicar-se entre elas, fazia-se necessário à criação de protocolos de comunicação padronizados. Tal padronização foi alcançada a partir da definição e implementação dos protocolos TCP/IP (*Transmission Control Protocol* - Controle de Transmissão de Protocolo/ *Internet Protocol* - Protocolo da Internet), que permitiu a comunicação entre computadores com arquiteturas díspares e também entre as redes. Era então criado o padrão de protocolo sobre o qual opera a Internet. (CASTELLS, 2004)

Diante disso, a padronização dos protocolos acarretou crescimento mais acelerado da rede de computadores, pois,

Em 1984, A National Science Foundation (NSF) dos EUA estabeleceu a sua própria rede informática de comunicações, NSFNET e, em 1988, começou a utilizar a ARPA-Internet como a sua espinha dorsal (back-bone)². Em fevereiro de 1990, a ARPANET, tecnologicamente obsoleta, foi desmontada. Posteriormente, tendo a Internet ficado livre do contexto militar, o Pentágono encarregou a Fundação Nacional para a Ciência da sua gestão. Mas o controle da rede por parte da NSF durou pouco. Com a tecnologia para a criação de redes informáticas aberta ao domínio público e com as telecomunicações em pleno processo de desregulação, a NSF procedeu imediatamente à privatização da Internet. (CASTELLS, 2004, p. 28)

Este momento configurou-se como ponto de partida para a real popularização da Internet e seu uso por parte de organizações do mundo corporativo, bem como da população em geral.

² Backbones são poderosos computadores conectados por linhas que têm a capacidade de dar vazão a grandes fluxos de dados, como canais de fibra óptica, elos de satélite e elos de transmissão por rádio. De acordo com Bogo (2000), traduzido para o português o termo backbone significa “coluna dorsal de uma rede” e representa a via principal de informações transferidas por uma rede, neste caso, a Internet.

Até então a Internet era utilizada apenas pela comunidade acadêmica e científica e a Net só existia no formato de texto e isto a limitava a um gueto de cientistas e centros de pesquisas.

A Internet tal qual hoje se configura, com sua interatividade, como arcabouço de redes interligadas de computadores e seus conteúdos multimídia, só se tornou possível pela contribuição do Cientista Tim Berners-Lee e ao, *Conseil Européen pour la Recherche Nucléaire* - Centro Europeu de Pesquisas Nucleares (CERN), que criaram a *World Wide Web (WWW)*, inicialmente interligando sistemas de pesquisas científicas e mais tarde acadêmicas, interligando Universidades. Assim, a rede coletiva ganhou uma maior divulgação pública a partir dos anos 90.

Em agosto de 1991, Tim Berners-Lee publicou seu novo projeto para a *World Wide Web*, dois anos depois de começar a criar o *Hypertext Markup Language* - Linguagem de Marcação de Hipertexto (HTML), o *Hyper Text Transfer Protocol* - Protocolo de Transferência de Hipertextos (HTTP) e as poucas primeiras páginas no CERN, na Suíça.

A partir do surgimento da *Web*, vários outros serviços foram desenvolvidos e disponibilizados na rede. Das versões modificadas da *WWW*, a que tinha uma aplicação mais comercial era a Mosaic, criada por Marc Andressen, do *National Center for Supercomputer Activit* - Centro Nacional para Atividade do Supercomputador NCSA, a qual tinha a capacidade gráfica avançada para obter e distribuir imagens através da Internet.

Esse software foi disponibilizado de forma gratuita na Internet, tal como a *Web*, e a partir de então outros *browsers* foram desenvolvidos. A filosofia de compartilhamento de saberes técnicos entre os cientistas e técnicos de informática no processo de construção da Internet permitiu o seu crescimento de forma bastante acelerada bem como a gratuidade dos serviços por ela prestados.

Muito embora esta tecnologia tenha iniciado o seu desenvolvimento no início dos anos 60, só a partir de 1995 é que ela surgiu para as pessoas, as empresas e para a sociedade em geral. Em 1996 a palavra Internet já era de uso

comum, principalmente nos países desenvolvidos, referindo-se na maioria das vezes a *www*.

A velocidade que a Internet impôs ao processo de comunicação entre as pessoas e as estruturas sociais, em geral, faz deste poderoso meio de comunicação uma importante ferramenta para o desenvolvimento científico e tecnológico, nos dias atuais. Para tanto, ainda existem pontos que devem ser melhorados para o uso cada vez mais proveitoso das informações que estão sendo disponibilizadas.

Assim, com o intuito de apresentar a configuração da internet no país, será apresentada a seguir uma abordagem concisa.

2.1.2 Internet no Brasil

No Brasil, a Internet também surgiu, num primeiro momento, no meio acadêmico e só depois foi que chegou ao acesso da população em geral. De acordo com Takahashi (2000), a história da Internet no Brasil se assemelha ao processo ocorrido nos EUA, em virtude do pioneirismo acadêmico e um posterior envolvimento governamental que alavancou o seu desenvolvimento.

A história da Internet no país começou tardiamente, comparando aos Estados Unidos, *locus* onde tudo começou. Em 1990, por uma iniciativa do Ministério de Ciência e Tecnologia (MCT) é criada a Rede Nacional de Pesquisa (RNP), cujo objetivo foi o de implantar uma infra-estrutura de serviços de Internet no país. Seu lançamento contou com o apoio de algumas fundações de pesquisa estaduais sob a coordenação do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Em 1991, com a expansão da rede, uma operação acadêmica subordinada ao MCT permitiu os primeiros serviços de Internet, por meio de pontos de presença em 21 estados implantados pela RNP, com velocidades de transmissão de dados ainda bastante baixas. Porém, algumas iniciativas neste sentido foram empreendidas no país anteriormente a este período.

Até 1995, a RNP provia acesso à Internet apenas a instituições voltadas à educação e pesquisa no Brasil, foi quando o Governo Federal definiu as regras gerais para a disponibilização desse serviço.

Entretanto, os primeiros passos dados no Brasil em direção à Internet deram-se no ano de 1987, quando a Fundação de Pesquisa do Estado de São Paulo (FAPESP) e o Laboratório Nacional de Computação Científica (LNCC), conectaram-se a instituições nos EUA. Esta conexão era feita através da Empresa Brasileira de Telecomunicações (EMBRATEL). Da mesma forma aconteceu a conexão feita por outras instituições nessa época, porém este modelo apontava para a necessidade de criação de um *backbone* nacional, sem o qual o desenvolvimento da internet brasileira seria inviável. (HISTÓRIA DA INTERNET BRASIL, 2006)

Em abril de 1995, o Ministério das Comunicações e o Ministério da Ciência e Tecnologia decidiram lançar um esforço comum de implantação de uma rede Internet global e integrada, abrangendo todo tipo de uso. Surgiu então o *backbone* nacional de uso misto, ou seja, agora além de acadêmico passou a ser também comercial, a partir de sua expansão e reconfiguração.

Nesse período é criado então o Comitê Gestor da Internet (CG) cujos objetivos consistiam basicamente em fomentar o desenvolvimento de serviços de Internet, coordenar a atribuição de endereços eletrônicos, registro de domínios e interconexão de espinhas dorsais (*backbones*) e por fim coletar, organizar e disseminar informações sobre serviços de Internet. (HISTÓRIA DA INTERNET NO BRASIL, 2006).

Todo o esforço convergido para colocar o Brasil na rede mundial de computadores foi uma das ações para inserir o país no contexto da Sociedade da Informação, a qual representa a grande remodelação social que ocorreu no mundo inteiro em virtude do nascimento e crescimento da Internet bem como a aceleração do desenvolvimento científico e tecnológico decorrentes do contexto pós-guerra.

A possibilidade de acesso incondicional da informação, no que diz respeito ao tempo e espaço, fizeram com que a Internet, bem como a *Web* fossem vislumbradas como uma panacéia para os problemas de desigualdades socio-

culturais tanto entre pessoas quanto entre as nações. Isso devido ao fato de que a informação passou a configurar-se como um bem econômico e, ainda, como insumo estratégico de todo e qualquer processo de desenvolvimento.

Nesse sentido, o ex-embaixador e Ministro de Estado da Ciência e Tecnologia do Brasil, Ronaldo Mota Sardenberg afirmou que

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. (TAKAHASHI, 2000, p. v)

Dentro desta perspectiva a Internet passa a configurar-se como uma importante ferramenta para os países, bem como as pessoas se inserirem neste novo contexto, devido a grande possibilidade de acesso a informação que o ambiente digital proporciona, bem como a possibilidade de construção do conhecimento coletivo, através da troca de saberes pela rede.

Essa tecnologia mudou antigos paradigmas sociais. Marcou o início de uma nova Era, ou seja, surgia então o conceito de Sociedade da Infomação onde, segundo Takahashi (2000. p. 59) “a informação, as telecomunicações, as novas mídias, a informática em geral e a indústria eletro-eletrônica passaram a ocupar lugar central no processo de acumulação de capital.”

Desse modo, o maior acesso à informação e comunicação poderá conduzir as sociedades e suas relações sociais de maneira mais democráticas, mas também poderá gerar uma nova forma de exclusão, acentuando as desigualdades e exclusões já existentes, tanto entre sociedades, como, no interior de cada uma, entre setores e regiões de maior e menor renda. (TAKAHASHI, 2000)

Porém, se de um lado as novas tecnologias conduz a vertentes de inclusão, como também de exclusão, torna-se mister refletir acerca do real poder de transformação das tecnologias, de um modo geral. Pois, as tecnologias por si sós não são sinônimos de desenvolvimento, pois para que este se consolide se faz necessário que o homem se utilize desta tecnologia de forma racional e intencional,

ou seja, a partir do seu conhecimento ele se apropria da tecnologia para potencializar as suas ações em prol do desenvolvimento.

Neste sentido torna-se primordial que se busque cada vez mais entender o processo através do qual se dá a comunicação entre o homem e a máquina no ambiente da Internet, indiscutivelmente, um grande repositório informacional. Sobre este ponto é que iremos abordar a seguir através da descrição da *Web* e sua estrutura.

Capítulo 3

A World Wide Web e seus precursores

3. WORLD WIDE WEB (WWW) E SEUS PRECURSORES

Como já foi mencionado anteriormente, a *Web* nasceu a partir de um projeto desenvolvido por Tim Berners-Lee no CERN, atual *European Laboratory for Particle Physics* (Laboratório Europeu de Física Aplicada). Porém, necessário se faz um breve relato sobre a abordagem histórica desse projeto revolucionário que se apropriou de idéias de estudiosos que o precederam.

Nesse sentido, Robert Cailliau apud Gromov (2006), ressalta que a história da *Web* está diretamente ligada, além da internet, ao desenvolvimento do hipertexto.

Para Lévy (1999, p. 55),

[...] a abordagem mais simples do hipertexto é descrevê-lo, em oposição a um texto linear, como um texto estruturado por nós (os elementos de informação, parágrafos, páginas, imagens, seqüências musicais, etc.) e por *links* entre esses nós, referências, notas, ponteiros, “botões” indicando a passagem de um nó para outro. [...] Mantendo ainda a definição de “texto em rede” ou de rede de documentação, uma biblioteca pode ser considerada como um hipertexto. Nesse caso, a ligação entre os volumes é mantida pelas remissões, as notas de pé de página, as citações e as bibliografias, fichários e catálogos constituem os instrumentos de navegação global da biblioteca.

Partindo dessa premissa, pode-se inferir que, o grande precursor das idéias que culminaram com a concretização da *Web* foi Vanevar Bush, com a publicação em 1945 do seu artigo “As we may think”. Com isso, Bush apresenta o projeto Memex, um dispositivo, que tinha como proposta criar um sistema de recuperação da informação, baseado na forma como o homem pensa, ou seja, através de associações de idéias.

O sistema idealizado por Bush pressupunha a criação de um imenso reservatório multimídia de documentos abrangendo ao mesmo tempo imagens, sons, e textos, dentro do qual dispositivos periféricos do sistema permitiriam transformar automaticamente a palavra em texto escrito. Nessa perspectiva, a massa de documentos do sistema seria miniaturizada através da microfilmagem e utilização da fita magnética, que acabavam de ser descobertas naquela época. Assim, as informações seriam acessadas através de uma tela de televisão munida de auto-falantes. (BUSH apud CAMPOS, 2001)

Ainda, de acordo com Lévy (1993) apud Campos (2001, p.17)

[...] Além dos acessos clássicos por indexação, um comando simples permitiria ao feliz proprietário do Memex criar ligações independentes de qualquer classificação hierárquica entre uma informação e uma outra. Uma vez estabelecida a conexão, cada vez que determinado item fosse visualizado, todos os outros que tivessem sido ligados a ele poderiam ser instantaneamente recuperados, através de um simples toque de um botão [...].

Após duas décadas, Ted Nelson cunhava a palavra hipertexto, a qual ele diz ser uma “forma não linear de escrita a qual ramifica e permite ao leitor uma melhor leitura numa tela interativa”. (GROMOV, 2006).

Nelson (1965) apud Campos (2001) definiu, pela primeira vez, o termo hipertexto como um texto ou material imagético interconectado num modo complexo que não pode ser representado no papel. Desse modo, pode ser constituído de um sumário ou mapa com seus itens e suas inter-relações. Foi também em meados dos anos 60 que Nelson iniciou o desenvolvimento de um sistema chamado Xanadu, que tinha como objetivo ser um “repositório universal de todas as publicações produzidas por qualquer pessoa no mundo”. O sistema então seria capaz de interconectar todas essas publicações através de *links* de forma que todos os documentos estariam disponíveis no mundo inteiro. Assim sendo, temos então a primeira idéia de desterritorialidade.

Nos anos 70, a idéia de hipertexto amadurece, porém ainda de forma restrita aos seus desenvolvedores. E a partir dos anos 80, com a evolução tecnológica, aliada ao aumento da capacidade de armazenamento e processamento dos computadores, foi possível que os sistemas de hipertextos pudessem ser construídos com toda a abrangência, como a elaboração de telas gráficas de alta resolução, fazendo com que os documentos fossem representados com toda a sua potencialidade, com figuras, gráficos e toda a riqueza topográfica das publicações impressas. (CAMPOS 2001)

Segundo Campos (2001), apesar de todo o desenvolvimento tecnológico e a crescente popularização dos hipertextos nos anos 90, o início do século XXI ainda não apresentou de forma

[...]“materializada a idéia de Ted Nelson, de um hipertexto universal, pois entendemos que Nelson pensava o hipertexto além de nós interligados a outros nós; ele entendia que as ligações transportavam conjuntos de informações que continham significado e que, para fazerem sentido, deveriam ser conectados através de uma dada organicidade. Pensamos que a *Web* em toda a sua extensão, ainda não pode ser considerada um hipertexto universal, pois muito esforço de profissionais de informação ainda merece ser empreendido, no sentido do tratamento das informações contidas na “rede das redes”[...].”(CAMPOS, 2001, p.19)

Apesar de ainda não existir uma estrutura de hipertexto tal qual foi idealizada por Ted Nelson, é fato que a estrutura gerada pela *Web* de Berners-Lee nos permite um modelo de uso das informações na rede através de hiperlinks, os quais se assemelham à ação de navegar. Com a estrutura tecnológica proporcionada pela Internet faz com que se perceba no ciberespaço um sistema semelhante ao Xanadu.

A partir do exposto infere-se que Berners-Lee, para criar a *Web*, apropriou-se de ideais de pesquisadores tais como Vannevar Bush e Ted Nelson. Pois, utilizando-se da plataforma tecnológica criada inicialmente pela ARPA, ou seja, a atual Internet, Berners-Lee desenvolveu o seu projeto, juntamente com a equipe do CERN. Este, por sua vez, consistiu numa rede de hiperlinks que viabilizou a ligação entre documentos que poderiam estar situados em qualquer lugar na rede de computadores de seu laboratório ou do mundo, através da internet e do conceito de Universal Resource Locator – Localizador de Recursos Universal (URL). (BAX, 2001)

Segundo Araújo (2003), Berners-Lee percebeu que o conceito de hipertexto poderia ser aplicado à Internet com a utilização de três tecnologias conhecidas dentre os seus acrônimos: URL, HTTP e HTML. A URL determina que cada página da *Web* tenha um endereço único. Enquanto que o HTTP é um protocolo para transferência de textos, gráficos, som, vídeo, entre outros pela Internet. E por fim o HTML, que descreve como os textos e gráficos devem ser exibidos.

Esta linguagem, HTML, denominada de Linguagem de Marcação ou ainda linguagem de marcas são as grandes responsáveis pelo acelerado crescimento da *Web*, bem como o estabelecimento de uma interface mais amigável

entre homem e máquina. E sobre sua origem e desenvolvimento que será dissertado a seguir.

3.1 LINGUAGENS DE MARCAÇÃO

Na atual sociedade, o avanço tecnológico tem exigido cada vez mais desse segmento, um padrão de trabalho em rede, em especial, no processo de geração, transferência e uso de produtos e serviços de informação. Pois,

A informação e o computador são parceiros antigos, mas a intensificação e democratização do seu uso, aliadas à abstração sempre crescente do nível de interação e troca de informações, criaram terreno propício para a origem das chamadas linguagens de marcação. Este fenômeno foi marcante na primeira década dos anos 90, com o aparecimento da *Web*. Estas linguagens permitem a construção de padrões públicos e abertos que estão sendo criados para se tentarem maiores avanços no tratamento da informação; elas minimizam o problema de transferência de um formato de representação para outro e liberam a informação das tecnologias de informação proprietárias. (BAX, 2001, p.32)

No ambiente informático, há várias décadas que se recorre à marcação para estruturar documentos. Segundo Bax (2001) todo sistema editor de textos tem de embutir, juntamente com o texto editado, marcas que forneçam indicações de como o texto deve ser apresentado ao usuário. Dependendo do formato que se visualiza o texto, essas marcas podem ser escondidas ou explícitas pelo usuário, que obterá seu documento no formato visual desejado. De acordo com Marchal (2000) apud Bax (2001), cada *software* editor ou compilador de textos possui seu próprio conjunto de códigos com significado apenas para aquele sistema, que deverá rodar em um determinado sistema operacional ou em uma máquina específica. Diz-se destes sistemas que eles promovem uma marcação procedimental do texto, cada código indicando o procedimento a ser seguido para a apresentação do texto ao usuário.

Ainda de acordo com Bax, existem também as chamadas linguagens baseadas em marcação descritiva. Estas usam marcas (ou *tags*) para qualificar cada objeto do texto, o que consiste num primeiro passo para transformá-los em informação tratável por computador. Uma marca ou *tag* é tudo o que não for considerado conteúdo em um documento. Elas indicam a função (o propósito) da

informação no documento, em vez de como ela deve ser apresentada, ou seja, sua aparência física. A idéia básica é a de que o conteúdo do documento deve estar separado do estilo usado em sua apresentação. Cabe, à aplicação que interpreta a linguagem de marcação, formatar o texto em tempo real e apresentá-lo ao usuário.

De um modo mais geral, ao analisar um documento sob a ótica das linguagens de marcação, estes se constituem de três elementos, claramente identificáveis, ou seja, conteúdo, estrutura e estilo (ou formatação). O conteúdo diz respeito à informação propriamente dita, a estrutura define como se dá a organização da informação, ou das idéias, no documento e o estilo estabelece o visual de apresentação das informações ao usuário. (BAX, 2001)

Desse modo, a utilização de padrões de marcação internacionais, os quais serão citados posteriormente, permitem a criação de documentos portáteis, isto é, documentos que não são dependentes de um determinado *software*, *hardware*, ou sistema operacional para ser visualizado. Assim sendo, podem ser interpretados por aplicações presentes nos mais diversos ambientes computacionais que reconheça o padrão usado na criação do documento.

Neste sentido, Araújo (2003) cita o *Standard Generalize Markup Language* (SGML) – Linguagem de Marcação Generalizada Padrão (tradução nossa) como sendo a linguagem de marcação da qual muitas outras se originaram, tais como o *Hyper Text Markup Language* – Linguagem de Marcação do Hipertexto (HTML), *eXtensible Markup Language* – Linguagem de Marcação Extensível (XML), *eXtensible Hyper Text Markup Language* (XHTML) – Linguagem de Marcação de Hipertexto Extensível, entre outras.

De acordo com Bax (2001, p. 34)

Standard Generalized Markup Language, ou simplesmente SGML, é uma (meta)linguagem criada há aproximadamente 30 anos como um esforço para se definir uma linguagem de marcas para a representação de informações em texto (Edwards, 1997). A linguagem foi reconhecida como um padrão ISO (8879) em 1986. SGML não é um conjunto predeterminado de marcas, e sim uma linguagem para se definirem quaisquer conjuntos de marcas, uma linguagem *autodescritiva*; cada documento SGML carrega consigo sua própria especificação formal, o *Data Type Document* – Documento do Tipo de Dados (DTD).

O DTD, segundo Araújo (2005) consiste num arquivo contendo as definições de cada uma das marcações.

A *Web* tradicional é baseada em documentos escritos em HTML, uma linguagem usada, para a criação das páginas que estão na rede. A linguagem HTML, na prática, tornou-se inicialmente, inviável para a representação da estrutura de documentos, uma vez que o seu desenvolvimento acabou se desviando do objetivo a que pretendia, ou seja, ser uma alternativa a SGML, que é uma linguagem fortemente estruturadora de documentos. Entretanto, a HTML transformou-se em uma linguagem apenas de formatação de documentos. (BAX, 2001)

Por outro lado, a SGML ainda constituía-se em uma linguagem complexa, ficando restrita a poucos especialistas. A resposta à SGML viria com a linguagem XML, uma metalinguagem desenvolvida em 1998 pela W3C com o objetivo de ser uma linguagem com a capacidade de descrever outras linguagens de marcação. Com esta característica, que a diferencia da HTML, permite-se a organização da informação em um formato estruturado, sem a existência de elementos predefinidos pelo fabricante da linguagem. Essa facilidade oportuniza para que usuários dessa linguagem criem sua própria linguagem de descrição documental a cada domínio específico que seja necessário. (SANTOS, NASCIMENTO, OLIVEIRA, 2005).

A partir do exposto, torna-se visível a razão pela qual o ambiente *Web* se popularizou de forma tão rápida pelo mundo. Muito embora essa popularização ainda se dê de forma elitizada, devido a fatores sócio-econômicos e culturais, bem como cognitivos, indiscutivelmente a rede mundial de computadores já faz parte da realidade de diversas pessoas, as quais encontram neste ambiente um novo espaço para socialização de saberes, negócios corporativos, educação, pesquisas, entre outras atividades. Porém, como foi mencionado anteriormente, a *Web* é um ambiente compartilhado, mas altamente desestruturado, gerido de forma descentralizada e em crescimento explosivo, se constituindo num verdadeiro desafio para seus gestores, em especial para os profissionais da informação.

Em resposta a este ambiente anárquico, vê-se surgir inúmeras iniciativas no sentido de se criar uma *Web* mais organizada dentre os quais podemos citar a Internet2, www2, e o mais audacioso de todos a *Web Semântica*.

Esta última constitui-se um projeto o qual está sendo encabeçado pelo W3C. Esta, por sua vez, é uma organização, hoje liderada pelo criador da *Web*, Tim Berners-Lee, e que conta com a colaboração de diversos pesquisadores, de diversas áreas e, ainda, de inúmeras instituições as quais possuem um objetivo comum que é contribuir para a criação de padrões para o desenvolvimento da rede.

O projeto da *Web Semântica* vislumbra transformar a *Web* num grande sistema de recuperação da informação (SRI). Dentro desta perspectiva algumas colocações serão feitas a cerca dos mecanismos de recuperação da informação na rede.

3.2 RECUPERAÇÃO DA INFORMAÇÃO NA WEB

A recuperação da Informação configura-se como um processo que se dá mediante a comunicação entre um sistema de informação e o usuário para isso fazendo-se necessária a utilização de uma linguagem que seja comum aos dois sujeitos da operação.

A relevância desse processo reside no fato de que a informação constitui-se hoje em insumo básico para geração do conhecimento e, conseqüentemente, para o capital humano, o qual ocupa lugar de destaque na atual sociedade da informação.

De acordo com Cintra et al (2002, p. 10),

a informação cumpre papel decisivo na mudança dos destinos da humanidade, uma vez que ela está diretamente ligada ao conhecimento e ao desenvolvimento de cada uma das áreas do saber, já que todo conhecimento começa por algum tipo de informação.

Assim sendo, necessário se faz promover cada vez mais o acesso a esta informação através da disponibilização e recuperação, sendo este último

aspecto mais complexo devido ao crescimento vertiginoso com que se produz a massa documental desde o surgimento da imprensa.

Retrocedendo um pouco no tempo, ao longo da história da humanidade, percebemos que o processo de comunicação foi um fator primordial para o seu desenvolvimento. A partir da escrita, quando o homem deu os primeiros passos em direção ao registro do conhecimento, este possibilitou que o mesmo fosse compartilhado, promovendo assim uma aceleração cada vez maior na construção do saber. Isto porque o conhecimento deu seus primeiros passos para tornar-se explícito através de sua publicidade, ou seja, tornou-se público e acessível.

Sobre isso Barreto (1998, p. 123) afirma que,

A publicidade do conhecimento produzido é uma condição necessária para sua validação e socialização, construindo, também, um ciclo constante e auto-regenerativo: conhecimento → publicidade → opinião pública → novo conhecimento.

Porém, o fato de um documento ser publicado e tornar-se inacessível em virtude da sua não recuperabilidade, pode estar ligado à forma como esta foi representada num determinado sistema de forma insatisfatória, ou aquele usuário que busca não possui habilidades necessárias para recuperá-la de forma eficaz.

Desse modo, percebe-se que o processo de descrição dos documentos também implicou na criação de padrões que nortearam e ainda norteiam este procedimento, dentre os quais podemos citar as regras de catalogação e as linguagens documentárias.

Assim sendo, podemos citar as Regras de Catalogação Anglo-Americanas – *Anglo-American Cataloguing Rules* (AACR), os vocabulários controlados, *thesaurus* entre outros, como ferramentas utilizadas já há algum tempo para descrever conteúdos informacionais com vistas a sua recuperação, seja de forma manual ou automatizada.

Porém, com o crescimento da *Web* enquanto repositório informacional transferiu-se também para outros profissionais uma problemática que antes cabia geralmente aos profissionais bibliotecários, cuja atuação limitava-se ao contexto da biblioteca. Já para as informações em meio digital, algumas adaptações se fazem

necessárias devido à ubiquidade e volume imensurável de informações, ao que abordaremos num outro momento.

Por enquanto, aborda-se a relevância e instrumentos de representação da informação, já consolidados como ferramentas básicas na atuação dos profissionais bibliotecários.

De todos os procedimentos realizados por profissionais da informação na representação de um documento com fins de recuperação parece ser o da representação do seu conteúdo informacional, pois várias são as nuances que devem ser observadas no que diz respeito à linguagem.

Apesar de existirem entre os povos barreiras lingüísticas decorrentes de diferentes idiomas e contextos sociais, estes influenciam nas formas de comunicação e, que, se não forem consideradas no momento da representação do conteúdo informacional de determinado documento, este pode tornar-se irrecuperável.

A padronização da linguagem, através da Linguagem Documentária (LD) é apontada na literatura como uma possível solução para esta problemática. Pois, de acordo com Cintra et al. (2002, p. 16)

A função da LD é tratar o conhecimento dispondo como informação. Compete às LDs transformar estoques de conhecimentos em informações adequadas aos diferentes segmentos sociais. [...] Durante algum tempo acreditou-se que a disponibilização dos estoques seria suficiente para sua socialização. Mas, atualmente, o fundamental é a existência de uma forma de organização que garanta o partilhamento. Essa organização é a LD”.

No tocante às bases de dados bibliográficas, Gardin apud Lucas (1999), propõe o termo metalinguagem para as LDs (no caso os Tesauros), os quais são utilizados para representar o conteúdo dos documentos. Segundo esses autores, as linguagens documentárias constituem-se em sistemas de símbolos.

Assim sendo, outro novo conceito se insere no contexto biblioteconômico a partir dos neologismos do mundo informático, que são os metadados. Segundo Codina e Rovira (2006, p. 301) “Os documentalistas estão familiarizados com o conceito de metadados desde muitos anos (décadas?,

séculos?) sob outras denominações anteriores: catálogos, documentos secundários, classificações, etc.”(tradução nossa)

Senso e Rosa Piñero (2003) reforçam esta idéia ao afirmarem que estes não se configuram como algo completamente novo no universo biblioteconômico. O termo foi cunhado por Jack Myers, na década de 60 para descrever conjuntos de dados. O primeiro conceito atribuído e, atualmente, ainda o mais usado, foi o de dado sobre dado, uma vez que proporcionavam a informação mínima necessária para identificar um recurso.

Desse modo, os autores consideraram a catalogação como sendo um processo de geração de metadados. Porém, tendo em vista que a maioria dos sistemas de metadados tem sido criada não somente por profissionais da informação, como também por informáticos, técnicos de sistemas, entre outros.

Nesse sentido, o conceito é utilizado como um termo neutro, permitindo eliminar possíveis prejuízos por parte de todos aqueles que são menos próximos ao mundo bibliotecário e que coloca a todos os grupos profissionais envolvidos em seu desenvolvimento em uma posição de igualdade (SENSO; ROSA PIÑERO, 2003) (tradução nossa).

Desse modo, os autores propõem como conceito de metadado, “[...] toda informação descritiva sobre o contexto, qualidade, condição ou características de um recurso, dado ou objeto que tem a finalidade de facilitar sua recuperação, autenticação, avaliação, preservação ou interoperabilidade”.(SENSO; ROSA PIÑERO, 2003, p. 98, tradução nossa)

Porém, para um melhor entendimento sobre o que vem a ser na prática um metadado, Takahashi (2000) apresenta o seguinte conceito,

Matadados são dados que descrevem outros dados, em uma definição sem maior rigor técnico. Por exemplo, em um formulário qualquer que tenha sido preenchido com os dados de um indivíduo, há tipicamente, em cada campo do formulário, uma informação explicitando que tipo de dado deve ser ali descrito: por exemplo, NOME COMPLETO, ENDEREÇO, etc. Essa informação é um metadado. (TAKAHASHI, 2000, p. 59)

Desse modo, podemos dizer que a recuperação da informação, em parte está condicionada aos padrões de metadados utilizados. Neste sentido

podemos afirmar que, no contexto bibliotecário, os primeiros padrões de metadados foram estabelecidos pelas Regras de Catalogação Anglo Americana (AACR), que mais tarde evoluiu para AACR2. Este, por sua vez, foi utilizado como base para o desenvolvimento do formato *Machine Readable Cataloguing* – Catalogação Legível por Máquina (MARC), que tornou possível a leitura dos metadados (catalográficos) pela máquina e assim surgindo os catálogos em linha e, ainda, a possibilidade do trabalho cooperativo de catalogação em rede entre bibliotecas.

Porém, quando se trata de recuperação da informação, o ciberespaço é demasiadamente mais complexo que o ambiente das tradicionais bibliotecas e, também, apresenta um elevado e desordenado ritmo de crescimento.

Sobre isso, Loureiro e Francelin (2006) afirmam que, em termos de bibliotecas e centros de informação, em geral, criamos sistemas, metodologias e ferramentas sofisticadas para o tratamento, armazenamento e recuperação da informação. Alguns desses procedimentos foram adaptados para o ambiente virtual e outros precisam ser criados, considerando a informação eletrônica sem um similar físico, ou seja, constituída puramente em meio virtual.

Sobre isso Marcondes (2005, p. 99) afirma que,

A primeira tentativa de dar conta da explosão informacional em que se transformou a *Web* foram os catálogos, como o Yahoo (o primeiro catálogo da *Web*), e os chamados mecanismos de busca, como AltaVista, Lycos, *WebCrawler*, etc, e mais recentemente, o Google. Enquanto em catálogos como o Yahoo, a descoberta, avaliação e descrição e inclusão dos recursos *Web* na base de dados é feita por profissionais de informação, os mecanismos de busca, para indexarem a *Web*, possuem programas que visitam página por página da *Web*, percorrem o texto de cada página, extraindo daí palavras-chave e armazenam numa base de dados estas palavras-chave, associadas ao URL da página. É sobre esta base de dados que os usuários fazem suas buscas nos “sites” dos mecanismos de busca. Naturalmente, por ser uma indexação automática com base em palavras isoladas, sem nenhum controle terminológico, efetuada em páginas sobre os mais variados assuntos, diferentes idiomas e totalmente desprovida de qualquer informação contextual, os resultados têm baixíssima precisão (Sneiderman, 1997). Estudos como o citado reforçam a dimensão do problema localização/identificação colocado pela Internet.

Como já foi mencionado anteriormente, tal problema da recuperação da informação deve-se em parte à Linguagem HTML, visto que esta, em sua estrutura, inicialmente não se constitui de metadados e tão somente de *tags* que especificam a forma de apresentação do documento. Assim sendo, Marcondes

afirma que, como tentativa de suprir essa falha na constituição das páginas em HTML e permitir uma indexação por parte dos *softwares* dos motores de busca, foram criadas então as *tags* META. Porém esta ainda não foi suficiente, pois não possibilitou uma melhor recuperação de documentos gerados em HTML a partir de bancos de dados. Estas também chamadas de dinâmicas, pois elas criam um documento eletrônico inédito a partir de informações que fornecemos, tal como acontece num processo de atendimento via Internet. Para melhor exemplificar o que viria a ser esse documento, Marcondes afirma também que

Exemplos desta situação podem ser vistos quando consultamos o “site” da Receita Federal para saber se a devolução do Imposto de Renda já foi depositada em nossa conta corrente. Quando informamos o CPF, o sistema de banco de dados emite automaticamente um documento em HTML só para nós.”

Para o autor, tais documentos constituem a *Web* profunda, ao que aos outros documentos denominou-se *Web* superficial.

Para Codina (2003) a *Web* “não indexável”, forma pela qual se refere à *Web* profunda consiste em documentos que não são indexados pelos motores de busca. Tal fenômeno, segundo o autor, pode ocorrer devido a três fatores. Primeiro são os formatos, pdf (documentos *Acrobat*) e doc (documentos *Word*), os quais são fortes candidatos a integrar a *Web* invisível devido ao fato de que estes se integram com menos facilidade ao navegador e por isso são fortes candidatos a configurarem como documentos “não indexáveis”. Outro motivo pelo qual alguns documentos não são recuperados pela *Web* se deve ao fato de fazerem parte de *sites* que exigem senha para acesso.

Assim sendo, a *Web* invisível segundo Marcondes e Codina, pode ser até 500 vezes maior que a Internet visível.

Porém Codina (2003) ressalta que no que concerne o acesso ao conhecimento, através da busca e obtenção da informação, não existe problema que os documentos mencionados por Marcondes (2005) continuem sendo invisíveis, visto que em sua maioria são de interesse particular. Porém, os documentos que são disponibilizados em formatos diferentes do HTML precisam estar bem indexados nas bases de dados para poderem ser recuperados. Outro ponto abordado pelo

autor é a grande diversidade de interfaces de busca existente em cada base de dados.

Sobre este último ponto, Codina (2003, p. 5) afirma que

Observa-se que as barreiras ao conhecimento são duas: o conhecimento das fontes e o domínio da interface de usuário de cada fonte. Com efeito, em primeiro lugar, para que o usuário possa beneficiar-se dos conteúdos de uma base de dados é necessário ao menos que saiba de sua existência. Porém, supondo que saiba de sua existência, então deverá ter habilidades de uso de tal base de dados e cada base de dados não somente apresenta uma interface aos usuários diferente, como também um conjunto de funções diferentes. (tradução nossa)

Felizmente, as fronteiras da *Web* invisível não fazem mais que retroceder. Pois hoje já é possível recuperar através de motores de busca, tal como o Google, documentos de diversos formatos dentre os quais podemos citar: Acrobat (pdf), Word (doc), Excel (xls), Power Point (ppt, pps), texto Enriquecido (rtf) entre outros. Já no que diz respeito às bases de dados, torna-se cada vez mais crescente os buscadores especializados, onde podemos citar o *Scirus for scientific information only* (*Scirus* apenas para informação científica), o qual indexa documentos de diversas bases de dados e também documentos HTML de sites de instituições científicas e acadêmicas.

Vimos que outro fator que dificulta sobremaneira a recuperação da informação no ambiente digital é a forma como os conteúdos são inseridos neste ambiente. Partimos do pressuposto de que as publicações na *Web* atingem níveis elevados de crescimento e, portanto não existem profissionais de informações suficientes para representá-los em sua totalidade. Desse modo, aponta-se como uma alternativa criar uma estrutura simples de metadados que permita que os próprios autores “descrevam seus documentos ao publicá-los eletronicamente “na fonte”, conforme uma proposta bem antiga da biblioteconomia, retomada por Weibel (1995)”. (MARCONDES, 2005, p. 101)

Neste sentido, é válido mencionar o padrão de metadados para descrição de documentos eletrônicos proposto pela Dublin Core Metadata Initiative (DCMI), o qual denomina-se Dublin Core (DC). Este padrão constitui-se de 15 elementos descritivos, os quais são: **Subject** (assunto); **Title** (título); **Creator** (autor, responsável pelo documento); **Publisher** (publicador, quem torna o documento

disponível na Internet); **Contributor** (outros colaboradores - por ex. editores, tradutores, etc); **Description** (descrição, resumo, sumário); **Date** (data de publicação); **Type** (tipo de recurso - homepage, romance, poesia, software, dicionário); **Format** (formato do arquivo que contém o documento eletrônico - texto, PDF, LaTeX, HTML, WORD ou outro); **Identifier** (geralmente o URL de um documento eletrônico); **Relation** (relacionamentos com outros documentos, por ex. versões); **Source** (fonte ou origem); **Language** (idioma do documento); **Coverage** (cobertura - espacial ou temporal, sobre que lugar, ou sobre que época o texto se refere); **Rights** (texto livre especificando qualquer restrição referente a direitos autorais). (MARCONDES, 2005)

De acordo com Souza et al *apud* Loureiro e Francelin (2006, p. 7) o padrão DC

é a catalogação do dado ou descrição do recurso eletrônico. A expectativa é que autores ou *WEBS*iters sem conhecimento de catalogação sejam capazes de usar o Dublin Core para descrição de recursos eletrônicos, tornando suas coleções mais visíveis pelos engenhos de busca e sistemas de recuperação.

Porém, neste sentido, o mais complexo projeto é apontado na literatura como sendo a *Web Semântica*, a qual como o próprio nome já sugere, trata-se de uma *Web* com significado. Sobre esse novo conceito este será abordado no próximo capítulo.

Capítulo 4

Web Semântica: “O estado da arte”

4 WEB SEMÂNTICA: “o estado da arte”

A *Web Semântica* é um dos projetos da *World Wide Web Consortium* (W3C). Esta organização foi fundada em outubro de 1994, por Tim Berners-Lee, desde então diretor da organização, no Laboratório de Ciência da Computação do Instituto de Tecnologia de Massachussets (MIT/LCS) em colaboração com o CERN, onde a *Web* foi originada. Desse processo, participaram ainda a *Defense Advanced Research Projects Agency* (DARPA) e a Comissão Europeia (*European Comission*).

Tim Berners-Lee juntamente com outros pesquisadores, criou a W3C como um consórcio industrial dedicado à construção de consensos sobre as tecnologias *Web*. Tal iniciativa tem como objetivo desenvolver tecnologias que sejam compatíveis umas com as outras e que permitam que qualquer *hardware* ou *software* usados para acessar a *Web* trabalhem juntos, ou seja, seu objetivo é promover a interoperabilidade da *Web* (W3C).

Segundo Berners-Lee (W3C..., 2005), a

W3C é onde o futuro da *Web* é construído. Seus membros trabalham juntos para delinear padrões tecnológicos que construirão a universalidade da *Web*. A W3C cria o poder da comunicação, do intercâmbio de informações e formula aplicações dinâmicas para qualquer um, em qualquer lugar, a qualquer hora, usando qualquer dispositivo. (W3C..., 2005,)³ (tradução nossa)

Para atingir os seus objetivos, este organismo conta com a participação de organizações localizadas por todo o mundo e de várias áreas do conhecimento. Estas organizações juntamente com a equipe de especialistas técnicos dedicados em tempo integral à instituição, convergem esforços para criar padrões tecnológicos que promovam o desenvolvimento da *Web* e sua interoperabilidade. Com esta desterritorialização que ela promove, pretende-se que seja cada vez maior o número de localidades que se beneficiam das novas tecnologias bem como sejam diminuídas as barreiras lingüísticas no que diz respeito à documentação que se produz neste campo de conhecimento, uma vez que o inglês ainda constitui-se o idioma mais utilizado neste aspecto.

Nesta perspectiva, surge como um dos projetos a *Web Semântica*, o qual visa promover uma melhor comunicação entre homem-máquina, ou seja,

³ Texto retirado da Internet

objetiva permitir entendimento entre a linguagem entendida pelo homem e a linguagem entendida pela máquina de forma que a máquina consiga entender semanticamente um comando dado pelo usuário do sistema, no caso a *Web*.

Pierre Levy, em conferência no mês de julho aqui no Brasil, mencionou que a partir do desenvolvimento de uma *Web Semântica* possibilitará resolver problemas sociais de nossa época, os quais decorrem de problemas científicos de gestão do conhecimento, o problema tecnológico da indexação pragmática e semântica do ciberespaço e o problema sócio-político de governança da inteligência coletiva. Segundo o filósofo estes problemas existem devido inexistência de instrumentos tecnológicos de observação, a diversidade das linguagens formais sem aplicação computacional e a presença de conhecimentos tácitos, difíceis de se guardar e compartilhar. (BIREME, 2006)

Numa visão mais técnica, segundo a W3C, a *Web Semântica* é a representação de dados na *Web*. Isto é, um esforço colaborativo e compartilhado e, ainda, liderado pela W3C com participação de um grande número de pesquisadores e indústrias parceiras.

Baseia-se na Estrutura da Descrição de Recursos - *Resource Description Framework* (RDF), os quais integram uma variedade de aplicações usando XML para a sintaxe e URI para o nome. Também são de total relevância para a *Web Semântica*, as ontologias, que permitem representar explicitamente a semântica dos dados. Através dessas ontologias é possível elaborar uma rede enorme de conhecimento humano, complementando o processamento da máquina e melhorando qualitativamente o nível de serviços na *Web*. (ARAÚJO, 2003)

O interesse da tecnologia da *Web Semântica* está focado em dois pontos: primeiro, os formatos comuns para intercâmbio de dados, onde na *Web* original só existe o intercâmbio de documentos. Segundo, na linguagem para registrar como os dados se relacionam com objetos do mundo real. Dessa forma, pretende-se “permitir uma pessoa, ou uma máquina, partir de uma base de dados, e então mover-se através de um infinito número de bases de dados as quais estão conectadas, não por fios, mas, por serem sobre a mesma coisa.” (W3C) (tradução nossa).

Sobre a definição proposta por Berners-Lee, bem como a da W3C, acima descritas, Codina (2003a) faz duas críticas: a primeira é quanto ao conceito proposto que, segundo o autor, o texto não diz nada e questiona, “o que significa dizer que alguma coisa seja a representação de dados na *World Wide Web*? Nada” e continua

“ o resto da suposta definição é pior. Abandona claramente a intenção de dizer o que é a *Web Semântica* e se limita a assinalar, entre outras coisas, que integra uma variedade de aplicações. A segunda é que a *Web Semântica* não existe. Não sabemos se a *Web Semântica* será realidade algum dia, mas hoje por hoje, não existe “ nem se espera”(ao menos de maneira iminente). Apesar disso, se deve reconhecer nela uma autêntica idéia-força, no sentido de que é uma idéia que já foi capaz de motivar muitas energias (e muitas ilusões) e que, sem dúvida não deixará de arrojar resultados durante os próximos anos, porque sem dúvida continuará mobilizando energias” (CODINA, 2003a, p. 10, tradução nossa)

O autor então propõe a definição de *Web Semântica* como sendo um conjunto de iniciativas, tecnológicas, em sua maior parte, destinada a criar uma futura *World Wide Web* na qual os computadores possam processar as informações, isto é, representá-las, encontrá-las, gestioná-las, como se os mesmos possuíssem inteligência.

Complementando o conceito de *Web Semântica*, Souza e Alvarenga (2004, p. 134) acrescentam que

O projeto da *Web Semântica*, em sua essência, é a criação e implantação de padrões (*standards*) tecnológicos para permitir este panorama, que não somente facilite as trocas de informações entre agentes pessoais, mas principalmente estabeleça uma língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

Dito de outra forma, a *Web Semântica* pretende fazer com que a máquina consiga entender semanticamente uma determinada busca, o que se torna possível, de acordo com os autores, através da utilização da RDF para descrever padrões de metadados e representados através das ontologias.

Assim sendo, para que os computadores possam funcionar de forma tão inteligente se faz necessário que tenham acesso a coleções estruturadas de informações (dados e metadados) e de conjuntos de regras de inferência que ajudem no processo de dedução automática. Desse modo torna-se possível que

seja administrado o raciocínio automatizado, ou seja, a representação do conhecimento (BERNERS-LEE *apud* DZIEKANIAK, 2004)

Para visualizar melhor como se dá e qual a importância do RDF e as ontologias para a *Web Semântica* algumas considerações serão feitas a seguir acerca da estrutura da *Web Semântica*.

4.1 ESTRUTURA DA WEB SEMÂNTICA

A literatura consultada acerca da *Web Semântica* aponta para a interdisciplinaridade que esta tecnologia pressupõe.

Segundo Araújo (2003), para viabilizar esta tecnologia, muitas áreas distintas de conhecimentos são envolvidas neste processo, dentre as quais podemos citar

a Inteligência Artificial, que dá lógica e impõe as regras, os bancos de dados, que interagem as informações, e a plataforma computacional, que é a própria Internet. As páginas da *Web* devem estar estruturadas para facilitar a explicação de seu significado. Para construir aplicações que envolvam a *Web Semântica*, Berners-Lee et al (2001) propuseram uma arquitetura em camadas. (ARAUJO, 2003, p. 31)

Uma primeira estrutura de arquitetura da *Web Semântica* apresentou-se conforme mostra a figura posterior:

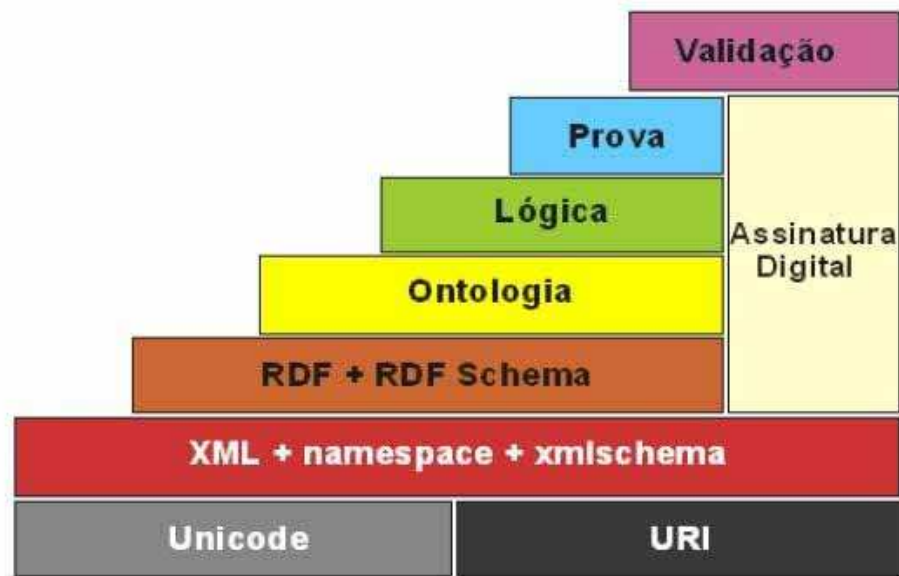


Figura 1 – Arquitetura proposta para a *Web Semântica*
Fonte: Carlan (2006)

Na primeira camada temos o Unicode⁴ que permite que textos e imagens possam ser lidos pelos computadores em qualquer lugar e o Identificador de Recursos Universal – *Universal Resource Identifiers* (URI), tecnologia através da qual as URL's são criadas, que fornece um endereço global único dos recursos disponibilizados na *Web*. Este poderá ser desde uma página na *Web*, um documento pessoal, um *videoclip*, um programa, dentre outros.

A próxima camada, onde se encontram as tecnologias XML, *Namespaces* e *XML Schema* permitem que os documentos da *Web* sejam estruturados em uma hierarquia de árvore, baseada em marcas, (*tags*), criadas pelo usuário. (ARAÚJO, 2003).

De acordo com Furgeri *apud* Farneda (2003, p 117), com a *XML Schema* é “possível não apenas especificar a sintaxe do documento XML como também especificar o número mínimo e máximo de vezes que um elemento pode ocorrer, criar listas e grupo de atributos.”

⁴ Unicode é um esquema de codificação de caracteres com 16 bits, incluindo não apenas os alfabetos gregos e romanos, mas também símbolos matemáticos, pontuação especial e conjuntos de caracteres não romanos. Trata-se de um padrão desenvolvido pelo Unicode Consortium e possui várias formas de representação tais como UTF-8, UTF-16 e UTF-32. (WIKIPÉDIA, 2006)

Porém, para Ferneda (2003), apesar de essas linguagens se constituírem num sólido alicerce, ela não faz parte da *Web Semântica*, ou seja, a definição da *Web Semântica* inicia-se de fato com a camada *RDF-RDF Schema*.

Para esse autor,

a semântica da linguagem XML é um subproduto da definição da estrutura de um documento. Portanto, a estrutura e a semântica se confundem no interior de um documento XML. A linguagem denominada RDF fornece um meio de agregar semântica a um documento sem se referir à sua estrutura. (FERNEDA, 2003, p. 122)

Sobre a RDF, Codina (2006) diz ser a resposta da engenharia da informática à necessidade de dispor de sistemas altamente homogêneos para descrever recursos digitais em particular e recursos de qualquer tipo. Desse modo, o RDF não é um conjunto de metadados predefinidos, como Dublin Core e sim um conjunto de especificações para criar qualquer tipo de conjunto de metadados seguindo um conjunto de normas pré-estabelecidas que tornem possível o intercâmbio de descrições criadas por agentes distintos e sem necessidade de que se tenham estabelecido qualquer acordo entre eles. (CODINA, 2006).

Outro conceito importante para a *Web Semântica* são as Ontologias, as quais são responsáveis pelo embasamento semântico desta tecnologia.

Segundo Ding et al (2005), sob uma perspectiva filosófica, a ontologia é a ciência que estuda a natureza do ser e sua existência. Etimologicamente o termo deriva das palavras gregas “onto”, que significa “ser”, e “logia”, que significa o discurso escrito ou falado. Já os cientistas da computação propõem uma nova interpretação das ontologias, que seria entendida como uma especificação de uma conceitualização.

Para os autores, tanto na Ciência da Computação como na Ciência da Informação, a reutilização do conhecimento é facilitada pelo uso de ontologias explícitas em oposição às ontologias implícitas, isto é, conhecimento codificado em sistemas de *software*. Por esta razão, apropriar-se de linguagens ontológicas são necessárias para compreender ontologias explícitas respeitando três aspectos importantes: conceitualização, vocabulário e axiomatização.

Segundo Araújo, (2003, p.14),

desde o início dos anos 90, as ontologias se tornaram um tópico comum de investigação na área de Inteligência Artificial, incluindo a engenharia e representação do conhecimento e o processamento da linguagem natural. Mas, recentemente a noção de ontologia tem-se expandido para as áreas de recuperação de informações na Internet, gestão do conhecimento e elaboração de sistemas educacionais inteligentes. A razão da sua popularidade está sustentada, basicamente, na promessa de que um determinado domínio de conhecimento pode ser representado computacionalmente, de modo que a comunicação entre pessoas e computadores se realiza automaticamente.

Neste sentido, vale citar o trabalho que Pierre Levy vem desenvolvendo há 15 anos, o qual se configura como um extenso mapeamento de conceitos na área das ciências humanas, visando o desenvolvimento de ferramentas que possam auxiliar as pessoas na busca por informações no ciberespaço. Trata-se da Meta Linguagem de Organização da Informação - *Information Economy Meta Language* (IEML) a qual está sendo desenvolvida pensando na sua utilização pela informática e por isso acredita-se que, ao mesmo tempo em que pode ser compreendida por cientistas sociais, pode igualmente ser traduzida por computadores. Por esse motivo acredita-se que será possível superar a diversidade de linguagens formais nas ciências humanas, inventando os meios que podem aproximar contextualmente seus diversos domínios. Todas as informações sobre este estudo encontram-se disponíveis no site www.ieml.org, o qual se apresenta no idioma francês, mas que segundo Lévy dentro de um prazo de aproximadamente três meses estará disponível em inglês. (LEVY, 2006)

Ainda sobre as Ontologias, Codina (2006, p. 304, tradução nossa) apresenta uma definição para o termo, porém deixando clara sua opinião cética a respeito, pois para o autor

As ontologias constituem o aspecto mais avançado (mais fantasioso?) da Web Semântica. Uma ontologia, no sentido da Web Semântica, é uma coleção coerente de enunciados sobre algum aspecto da realidade. Estes enunciados incluem declarações do tipo <um caminhão é um tipo de automóvel> e do tipo <um automóvel é um meio de transporte>; mas também restrições como a seguinte: <um caminhão não serve para transportar passageiros>, ou como a seguinte: <um caminhão não serve para o transporte marítimo>, etc.

Segundo Ramalho (2006), no contexto da *Web Semântica*, várias linguagens computacionais foram criadas especificamente para o desenvolvimento

de ontologias, dentre elas: *Ontology eXange Language* (XOL), *Ontology Markup Language* (OML), *Ontology Inference Layer* (OIL) e *Darpa Agent Markup Language* (DAML). A união dessas duas últimas originou a *Ontological Web Language* (OWL) a qual se constitui o padrão recomendado pela W3C para construção de ontologias. De acordo com o autor, esta permite descrever formalmente os aspectos semânticos dos termos utilizados e seus respectivos relacionamentos, possibilitando representações mais abrangentes das linguagens RDF e *RDF Schema*, o que favorece uma maior interoperabilidade.

As demais camadas que estruturam a arquitetura da *Web Semântica* ainda se encontram em processo de desenvolvimento (ARAÚJO, 2003; RAMALHO, 2006). Porém, algumas considerações sobre as funções que estas camadas deverão desempenhar são feitas pelos autores que abordam este tema.

De forma resumida, Araújo (2003) afirma que a camada de lógica tem como objetivo especificar linguagens de lógica mais poderosas que as atuais, para facilitar a construção de inferências. Já a camada de “prova” seria a especificação de uma linguagem para provar que as informações trocadas entre agentes e máquinas são verdadeiras. Porém a troca pode ser verdadeira, mas a informação pode ser falsa. E a camada “validação” é para verificar, através de assinatura digital, se a pessoa que realiza as transações previstas pelas aplicações é, de fato, “ela mesma”.

Já para Ramalho (2006), esta última camada espera garantir que as informações estejam representadas de modo correto, possibilitando certo grau de confiabilidade. E não apenas a certificação de sua origem quanto a quem produziu.

Ramalho acrescenta que a arquitetura da *Web Semântica*, desde 2000, quando Berners-Lee propôs o primeiro modelo, tal qual está representado na Figura 1, este já sofreu algumas modificações em virtude dos avanços alcançados através das pesquisas, principalmente em nível de ontologias. Segundo o autor, três modelos diferentes já foram propostos, onde outros padrões foram incorporados à estrutura, sendo que em 2005 foram feitas recomendações de mudanças às quais se encontram representadas estruturalmente na Figura 2.

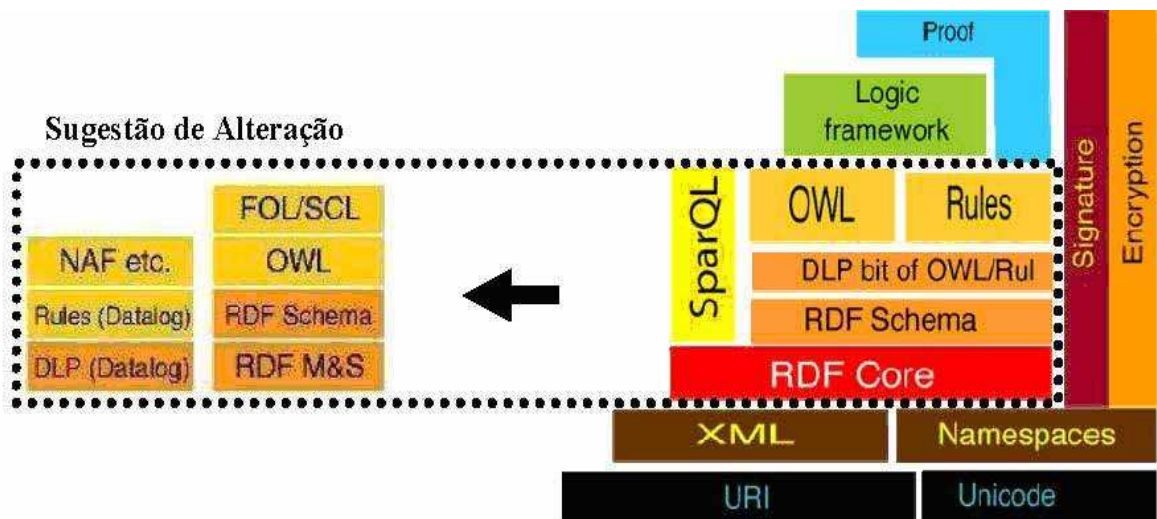


Figura 2 – Sugestão de alteração na Arquitetura proposta para a Web Semântica
Fonte: Ramalho (2006)

De acordo com Ramalho (2006) sem levar em consideração as modificações, a estrutura representada na figura 2, representa o último modelo proposto por Berners-Lee em 2005 para a *Web Semântica*. O autor ainda prevê que muitas alterações ainda deverão ser feitas visto que ainda não existem tecnologias recomendadas para as camadas “*Logic Framework*” e “*Proof*” (respectivamente “estrutura lógica” e “prova”).

Porém, apesar das mudanças em cada uma das camadas advindas de críticas e sugestões, vale ressaltar que em sua essência, os conceitos básicos parecem permanecer estáveis, os quais são: uma camada estrutural, camada sintática, camada semântica, camada lógica e por último a camada confiança. (RAMALHO, 2006)

A partir do exposto, foi possível visualizar como a *Web Semântica* pretende alcançar os objetivos propostos em seu projeto. Entretanto, os desafios persistem e pelo exposto na literatura, muitos esforços de profissionais e especialistas ainda deverão ser empreendidos até que se chegue a um tal nível de satisfação no processo de recuperação da informação.

A literatura em geral que aborda essa temática trata a *Web semântica* de forma bastante otimista, porém o espanhol Codina (2003) não acredita que a

tecnologia da *Web Semântica* possa, a médio ou curto prazo, alcançar seus propósitos, muito embora reconheça que os estudos podem levar ao desenvolvimento de áreas correlatas.

Segundo o autor

o objetivo da *Web semântica* é magnífico, produzirá importantes avanços em alguns ou em todos os terrenos relacionados com a representação e o acesso ao conhecimento e todos devemos apoiá-lo. Mas,[...], haveria que evitar retornar a irracionalidade dos primeiros anos da *Web*. Foram uns anos de chumbo no que se refere ao pensamento crítico: não havia dia que alguém não anunciasse uma suposta lei histórica, econômica, social e política que a Internet não rompeu. Isso produziu, entre outras coisas, a borbulha da Internet, muita especulação e muitos recursos atirados pela janela. (CODINA, 2003, p. 6) (tradução nossa)

Sobre o estado atual da *Web Semântica*, esta ainda não existe como no formato de um motor de busca, porém as tecnologias propostas para a sua existência já vêm sendo utilizadas por alguns organismos e empresas que atuam no comércio eletrônico, bem como os programas de educação à distância, as bibliotecas digitais, entre outras.

Para Shadbolt, Hall e Berners-Lee (2006), a idéia de criação de uma *Web* de dados, ao invés de documentos, foi idealizada em 1994, quando o seu idealizador, Berners-Lee, fez sua primeira explanação sobre a *Web Semântica*, na primeira Conferência da *World Wide Web*.

Porém, esta só foi apresentada ao público através de um artigo publicado na *Scientific American* em 2001. Entretanto, após nove anos a mesma ainda não se tornou realidade. Muito embora os avanços sejam notórios no que diz respeito ao desenvolvimento das linguagens utilizadas na *Web*, as quais permitem um uso cada vez maior e mais proveitoso da informação neste ambiente.

Ainda de acordo com os autores, cada vez mais cresce a necessidade de promover a integração dos dados de forma que se torne possível à construção de uma *e-ciência*, ou seja, a integração dos vários saberes correlatos de forma que o desenvolvimento dos mesmos se dê de forma sistêmica.

Para os autores, várias especialidades de áreas do conhecimento desenvolveram suas ontologias, porém o grande desafio agora é o de integrá-las através do seu compartilhamento, como por exemplo:

a ciência ambiental procura integrar dados de hidrobiologia, climatologia, ecologia e oceanografia. A necessidade de entender sistemas através de largas escalas de distribuição é evidente em qualquer ciência e apresenta uma pressão no requerimento de integração de dados e informação. (SHADBOLT, HALL e BERNERS-LEE, 2006, p. 97, tradução nossa)

Os autores mencionam ainda as ações governamentais e comerciais no sentido de financiar o projeto da *Web Semântica*, como por exemplo, o Reino Unido tem desenvolvido um Vocabulário Integrado do Setor Público (*Integrated Public Sector Vocabulary*), mas que ainda assim este ainda não se encontra acessível às pessoas.

Os progressos se deram a nível tecnológico no que diz respeito às URI's, RDF e RDFS, OWL entre outras. Porém as ontologias é que se configuram como o real desafio para a *Web semântica*, pois depende de métodos eficazes de desenvolvimento e gerenciamento, visto que, por tratar-se de uma linguagem, esta se apropria de características que são próprias da mesma, tal como a sua necessidade de atualização, própria do seu desenvolvimento.

Porém os pontos críticos da *Web Semântica* são identificados pelos autores através dos seguintes questionamentos:

[...].Como poderemos efetivamente perguntar a grandes números de descentralizados repositórios informacionais de escalas variadas? Como poderemos alinhar e mapear entre as ontologias? Como construiremos um *browser* (navegador) na *Web Semântica* que efetivamente visualize e navegue nos enormes gráficos RDF conectados? Como estabelecer verdade e autoridade aos conteúdos? (SHADBOLT, HALL e BERNERS-LEE, 2006, p. 100, tradução nossa)

Diante disso, os mesmos concluem que:

Nós precisaremos da ajuda de pesquisadores em áreas tão diversas quanto à análise da rede social e a epidemiologia para entender como informação e os conceitos se espalham na *Web* e como estabelecer suas autoridades e veracidades. Nós não devemos perder de vista o fato de que a *Web*, e certamente muitos dos nossos importantes ambientes digitais dependem fundamentalmente de determinadas suposições gerais sobre o comportamento humano. (SHADBOLT, HALL e BERNERS-LEE, 2006, p. 100 tradução nossa)

A expectativa dos autores é de que as tecnologias desenvolvidas com vistas à criação e crescimento da *Web Semântica* contribuam também para o crescimento da *Web* ciência, ou seja, uma ciência que busque o desenvolvimento, desdobramento e entendimento da distribuição de sistemas informacionais, sistemas de humanos e máquinas, operando numa escala global. (BERNERS-LEE, SHADBOLT, HALL, 2006)

A partir do exposto percebe-se que o estudo da *Web Semântica* envolve apropriação de alguns conceitos próprios da área de informática, bem como de outras áreas correlatas. Portanto, para a elaboração do seu “estado da arte” se fez necessária à adoção de uma metodologia que norteasse de forma adequada a elaboração desse tipo de estudo, sobre a qual será abordada no capítulo que se segue.

Capítulo 5

Metodologia

5 METODOLOGIA

O presente estudo caracteriza-se como uma revisão de literatura acerca da *Web Semântica*, sobre o qual se tornou possível contextualizar este objeto de pesquisa a partir do Ciberespaço e pautado inicialmente na pesquisa bibliográfica e discussão dos textos.

Em seguida, o processo de leitura se deu mediante o embasamento necessário para a análise e síntese que constituíram o “estado da arte” do tema em discussão.

No intuito de construir um conhecimento teórico sobre a *Web Semântica*, o presente trabalho caracteriza-se como uma pesquisa de revisão bibliográfica, na qual se buscou no referencial teórico atinente às principais questões estabelecidas no objetivo deste trabalho, para que pudessem ser localizadas as contribuições científicas sobre esse assunto.

No que diz respeito à metodologia utilizada, Gil *apud* Silva e Menezes (2001, p. 21) afirma que uma pesquisa exploratória “visa proporcionar maior familiaridade com o problema a fim de torná-lo explícito ou a construir hipóteses.” Assim sendo, envolve levantamento bibliográfico, e assume em geral as formas de pesquisas bibliográficas.

Para o mesmo autor, a pesquisa bibliográfica classifica uma pesquisa quanto ao procedimento técnico, ou seja, esta se constitui em uma “pesquisa elaborada a partir de material já publicado, constituído principalmente de livros, artigos de periódicos e atualmente com material disponibilizado na Internet.” (GIL *apud* SILVA e MENEZES, 2001, P. 21).

Por tratar-se de um trabalho de revisão de literatura, convém destacar o que afirmam Silva e Menezes (2001) que o mesmo deverá responder às seguintes questões: quem já escreveu e o que já foi publicado sobre o assunto, que aspectos já foram abordados, quais as lacunas existentes na literatura?. Deste modo, dentre os vários objetivos que poderão configurar-se neste tipo de estudo, um deles é determinar o “estado da arte”, no qual o pesquisador procura mostrar através da literatura existente, “o que já se sabe sobre o tema, quais as lacunas existentes e

onde se encontram os principais entraves teóricos ou metodológicos.” (LUNA *apud* SILVA e MENEZES, 2001, p. 37)

Os procedimentos adotados neste estudo buscaram mapear e analisar textos que abordassem a *Web Semântica* sob a ótica de profissionais de áreas distintas, mas que interdisciplinarmente contribuem para a área, bem como a reconhecem como uma temática a ser explorada. Assim sendo, o período de abrangência das obras selecionadas datam do ano de 2001, ano em que foi publicado o primeiro artigo sobre a *Web Semântica* até os dias atuais. Por tratar-se de um assunto relativamente atual a sua delimitação cronológica tornou-se *sine qua non*. Desse modo, o total de textos analisados foi de 30. documentos.

A metodologia adotada permitiu abordar os aspectos mencionados nos objetivos específicos definidos anteriormente, o que proporcionou a compreensão e concretização do seu objetivo geral a partir do inventário da produção científica nesse campo do saber.

Pois, segundo Ferreira (2002), um estudo que se caracteriza como “estado da arte” constitui-se em dois momentos distintos, sendo o primeiro no qual o pesquisador

“interage com a produção acadêmica através da quantificação e de identificação de dados bibliográficos, com o objetivo de mapear essa produção num período delimitado, em ano, locais, áreas de produção. [...] Um segundo momento é aquele em que o pesquisador se pergunta sobre a possibilidade de inventariar essa produção, imaginando tendências, ênfases, escolhas metodológicas e teóricas, aproximando ou diferenciando trabalhos entre si, na escrita de uma história de uma determinada área do conhecimento”. (FERREIRA, 2002, p. 265)

Assim sendo, os procedimentos metodológicos adotados seguiram os seguintes passos: 1) Levantamento bibliográfico: realizado em nível nacional e internacional em fontes bibliográficas primárias (livros, periódicos, anais de congresso, dissertações, teses e documentos eletrônicos da Internet), e também secundárias (base de dados textuais e referenciais como: Scielo, Periódicos Capes) de áreas correlatas. 2). Seleção dos documentos e os critérios de pertinência quanto aos assuntos presentes na pesquisa. 3. Após o levantamento bibliográfico e seleção dos materiais foram realizadas as leituras analíticas dos textos selecionados. Estes, por sua vez, além de demandarem a leitura de outros assuntos correlatos para seu melhor entendimento, proporcionaram a consolidação de uma base teórica para um maior entendimento e definição do problema de pesquisa e, ainda, dos processos de

tratamento da informação que viabilizaram a base tecnológica para o surgimento do Ciberespaço e da *Web Semântica*, bem como sua contribuição no processo de produção, busca e recuperação da informação no ambiente digital.

Observou-se que a literatura nacional sobre o assunto encontra-se ainda em processo de construção, os quais se apresentam em sua maioria na forma de trabalhos monográficos, tais como teses e dissertações, bem como artigos de periódicos e trabalhos de eventos. Portanto, a leitura de documentos em outros idiomas tais como inglês e espanhol se fizeram necessários no transcorrer desse trabalho.

Capítulo 6

Considerações Finais

6 CONSIDERAÇÕES FINAIS

Durante o desenvolvimento desta pesquisa foi possível identificar que a *Web Semântica* é uma área do saber relativamente nova e por isso a sua literatura ainda é incipiente, principalmente no país. Os documentos acerca do tema apresentam-se na sua maioria em forma de trabalhos monográficos, teses e dissertações, bem como artigos de periódicos e trabalhos de eventos.

Vale salientar que as principais fontes encontram-se nos idiomas inglês e espanhol, o que por vezes configura-se como uma barreira no desenvolvimento de estudos dessa natureza, caso o pesquisador não domine esses idiomas.

Como referência primeira sobre o tema constitui-se o próprio portal do organismo que a projetou, ou seja, a W3C, onde pode-se encontrar *links* que remetem a outros *sites* de assuntos correlatos, alguns artigos publicados sobre a *Web Semântica*, inclusive o primeiro artigo publicado na *Scientific American* onde Tim Berners-Lee apresenta pela primeira vez ao grande público a *Web Semântica* e ainda listas de perguntas freqüentes dos usuários.

O fato de a *Web Semântica* objetivar o processo de recuperação da informação no contexto da *Web* torna-se necessário que cada vez mais profissionais bibliotecários passem a investigá-la de forma a não só contribuir para o seu processo de desenvolvimento, participando de grupos interdisciplinares de pesquisa, como também de produzir uma literatura nacional acerca da temática que torne facilitado os estudos desenvolvidos nesse novo e crescente campo do saber.

Sobre a sua aplicabilidade, são inúmeras as possibilidades, sendo que no contexto biblioteconômico vale ressaltar as bibliotecas digitais, as quais poderão beneficiar-se sobremaneira dessa nova tecnologia.

Portanto, dentre os principais resultados desta pesquisa, vale ressaltar a oportunidade que a autora teve de se familiarizar com a temática, o que possibilitou uma identificação e, conseqüentemente, um grande aprendizado, a partir da explicitação e evolução da *Web Semântica*, por meio do estado da arte.

Referências

REFERÊNCIAS

ABADAL, E.; CODINA, L. Recuperación de información. En: **Bases de datos documentales**: Características, funciones y método. Capítulo 2. Madrid: Síntesis, 2005, p. 29-92. Disponível em: < <http://www.lluiscodina.com/>>. Acesso em: 12 out. 2006.

ALMEIDA, Maurício B.; BAX, Marcello P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção **Ci. Inf.**, Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003.

ALVES, Rachel. C. V. **Web Semântica**: uma análise focada no uso de metadados. 2005. 180 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia Ciências, Universidade Estadual Paulista, Marília, 2005.

ARAÚJO, Moysés. **Educação a distância e a web semântica**: modelagem ontológica de materiais e objetos de aprendizagem para a plataforma COL. 2003. 191 f. Tese (Doutorado em Engenharia de Sistemas Digitais) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2003.

BARRETO, Aldo Albuquerque. A condição da informação. **Revista São Paulo em perspectiva**, São Paulo, v. 16, n. 3, p. 67-74, 2002. Disponível em: <http://aldoibct.bighost.com.br/condicao.htm#_edn19>. Acesso em: 21 jul. 2006.

BARRETO, Aldo Albuquerque. Mudança estrutural no fluxo do conhecimento: a comunicação eletrônica. **Ci. Inf.**, Brasília, v. 27, n. 2, p. 122-127, maio/ago. 1998.

BARRETO, Aldo Albuquerque. Os destinos da Ciência da Informação: entre o cristal e a chama. **Datagramazero**, n. 0, p. 1-8, dez. 1999. disponível em: <http://www.dgz.org.br/dez/F_I_aut.htm>. Acesso em: 20 maio 2005.

BAX, Marcello Peixoto. Introdução às linguagens de marcas. **Ci. Inf.**, Brasília, v. 30, n. 1, P. 32-38, 2001. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100005&lng=pt&nrm=iso>. Acesso em: 29 ago 2006.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a evolution of new possibilities. **Scientific American**, v. 284, n. 55, p. 34-43, may 2001.

BIREME. Pierre Lévy faz conferência sobre Web Semântica em evento da Rede Unida. **Newsletter Bvs**, n. 053, 21 jul. 2006. Disponível em: <<http://espacio.bvsalud.org/boletim.php?newsletter=20060721&newsLang=pt&newsName=Newsletter%20BVS%20053%2021/julho/2006&articleId=07141609200607>>. Acesso em: 13 out. 2006.

BOGO, Kellen Cristina. A história da Internet: como tudo começou. **Kplus cosmo**. Disponível em: <<http://kplus.cosmo.com.br/materia.asp?co=11&rv=Vivencia>>. Acesso em: 21 jul. 2006. (Matéria publicada em 2000)

CAMPOS, M. L. M.; CAMPOS, M. L. A. ; CAMPOS, L. M. *Web semântica e a gestão de conteúdos informacionais*. In.: MARCONDES, Carlos H. et al. (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005. p. 55-75.

CAMPOS, Maria Luiza de A. **A organização de unidades do conhecimento em hiperdocumentos: o modelo conceitual como um espaço comunicacional para realização da autoria**. 2001. 198 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001.

CARLAN, Eliana. **Ontologia e Web semântica**. 2006. 61f. Monografia (Bacharelado em Biblioteconomia) - Faculdade de Economia, Administração, Contabilidade e Ciência da Documentação e Informação, Universidade de Brasília, Brasília, 2006.

CASTELLS, Manuel. **A galáxia da Internet: reflexões sobre Internet, negócios e sociedade**. Lisboa: Fundação Calouste Gulbenkian, 2004. 325 p. Tradução de Rita Espanha.

CASTELLS, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999. (A era da informação: economia, sociedade e cultura, v.1)

CESARINO, Maria A. da N. Sistemas de recuperação da informação. **R. Esc. Bibliotecon. UFMG**, Belo Horizonte, v. 14, n. 2, p. 157-168, set. 1985.

CINTRA, Anna Maria Marques et al. **Para entender as linguagens documentárias**. 2. ed. São Paulo: Polis, 2002. 96 p.

CODINA, Lluís. Internet invisible y *Web semântica*: el futuro de los sistemas de información en línea?. **Revista tradumática**: traducció i tecnologies de la informació i la comunicació, [S. l.], n. 2, p. 1-16, nov. 2003. Disponível em: <<http://www.ddd.uab.es/pub/tradumatica/15787559n2a6.pdf>>. Acesso em: 20 set. 2006.

CODINA, Lluís. La Web semântica: una vision critica. **Profesional de la Informacion**, v. 12, n. 2, , pp. 149-152. March/April 2003

CODINA, Lluís; ROVIRA, Cristòfol. Recursos sobre la web semântica. **Revista Española de Documentación Científica**, [S. l.], v. 29, n. 2, p. 297-305, abr./jun. 2006.

CUNHA, Isabel M. R. Ferin; KOBASHI, Nair Yumiko; AMARO, Regina Keiko Obata F. Revisão bibliográfica. In.: SMIT, Johanna W. (Coord.). **Análise documentária: a análise da síntese**. Brasília: IBICT, 1987. p. 114-133.

DIAS, Guilherme Ataíde ; ARAÚJO, Eliany Alvarenga de ; SILVA, José Wendell de Moraes ; ALMEIDA, Marcelo José Siqueira Coutinho de . Indexando e Recuperando Informações de um Periódico Científico Eletrônico Especificadas em Metadados Dublin Core Conforme o Modelo de Dados RDF. In: SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 2., 2005, São Paulo, 2005. **Anais...** . São Paulo: [s. n.], 2005.

DING, Li et al. **Using ontologies in the semantic web: a survey**. TR-CS, Baltimore, 2005. Disponível em: <<http://ebiquity.umbc.edu/paper/html/id/266/Using-Ontologies-in-the-Semantic-Web-A-Survey>> . Acesso em 20 jun. 2006.

DZIEKANIAK, Gisele Vasconcelos; KIRINUS, Josiane Boeira. Web Semântica. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.**, Florianópolis, n. 18, p. 20-39, 2º sem. 2004

FERNEDA, Edberto. Recuperação da informação: análise sobre a contribuição da ciência da computação para a ciência da Informação. 2003. 147 f. Tese (Doutorado em Ciências da Comunicação) Escola de Comunicação e Artes, universidade de São Paulo, 2003

FERREIRA, Aurélio Buarque de Holanda. Internet. In: FERREIRA, Aurélio Buarque de Holanda. **Miniaurélio século xxi escolar**: o minidicionário da língua portuguesa. 4. ed. Rio de Janeiro: Nova Fronteira, 2001. p. 397.

FERREIRA, Norma Sandra de Almeida. As pesquisas denominadas “estado da arte”. **Educação e Sociedade**, ano xxiii, n. 79, p. 257-272, ago. 2002.

GARCIA-JIMÉNEZ, Antonio. Instrumentos de representación del conocimiento: tesauros versus ontologias. **Anales de documentación**, n. 7, p. 79-95, 2004.

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

GROMOV, Gregory R.. **History of Internet and WWW**: The Roads and Crossroads. Disponível em: <<http://www.netvalley.com/intval3.html>> . Acesso em: 10 jun. 2006.

HISTÓRIA DA INTERNET NO BRASIL. Disponível em : <<http://homepages.dcc.ufmg.br/~mlbc/cursos/internet/historia/Brasil.html>> . Acesso em: 15 out 2006.

LESK, Michael. The Seven Ages of Information Retrieval. **Udt Occasional Papers**, Canada, v. 5, n., p.1-18, mar. 1996. Disponível em: <www.ifla.org/VI/5/op/udtop5/udtop5.htm - 61k>. Acesso em: 20 out. 2006.

LEVY, Pierre. **Cibercultura**. São Paulo: Ed. 34, 1999. 264 p. (Coleção Trans). Tradução de Carlos Irineu da Costa.

LEVY, Pierre. **Semantic Web** [mensagem pessoal]. Mensagem recebida por: <jacquelinecunh@gmail.com> em 27 nov. 2006.

LOUREIRO, Mônica de Fátima; FRANCELIN, Marivalde Moacir. Metadados e sistema de recuperação de informação on line:: uma questão de semântica?. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 14., 2006, Bahia. **Anais...** . Bahia: Ufba, 2006. p. 1 - 15. CD-ROM.

LUCAS, Clarinda Rodrigues. A metalinguagem como lugar da interpretação: terminologia e bases de dados informatizadas. **DELTA**, São Paulo, v. 15, n. 1, 1999. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttex&pid=s0102-44501999000100007&ing=pt&nrm=iso>. Acesso em: 02 out. 2006.

MARCONDES, Carlos Henrique. Metadados: descrição e recuperação de informações na Web. In.: _____. et al. (orgs.). **Bibliotecas digitais: saberes e práticas**. Salvador/Brasília: UFBA/IBICT, 2005. p. 97-103.

MATTOS, Fernando Augusto. Exclusão digital e social: elementos para uma discussão. **Transinformação**, Campinas, Edição Especial, n. 15, p.91-115, set./dez. 2003. Quadrimestral. Disponível em: <<http://revistas.puc-campinas.edu.br/transinfo>>. Acesso em: 20 abr. 2006.

MÉNDEZ, Eva; SENSO, José A.. **Introducción a los metadatos**: estándares y aplicación. Disponível em: <<http://www.sedic.es/autoformacion/metadatos/introduccion.htm>>. Acesso em: 10 ago. 2006.

MICHAELIS. **Dicionário Michaelis**: português, inglês, espanhol. São Paulo: Melhoramentos, 2000. CD-ROM.

PÉREZ-GÓMEZ, Assunción. Ontological Engineering: a state of the art. Expert update. [S.l.], n.2, v. 3, p.33–43, Autumn 1999. Disponível em:<<http://citeseer.ist.psu.edu/444416.html>>. Acesso em: 20 out. 2006.

PIMENTA, Márcia Teresa da Rocha. O profissional da informação e as novas mediações no atendimento ao cliente. In.: CASTRO, César Augusto (Org.). **Ciência da Informação e Biblioteconomia**: Múltiplos discursos. São Luís: EDUFMA, 2002. p. 121 - 142.

RAMALHO, Rogério Aparecido Sá. **Web Semântica** : aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. 2006 . 120 f. Dissertação (Mestrado em Ciência da informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2006.

RIBEIRO, Cristina. As ciências documentais e a construção da web semântica. **Biblioteca Digital**, Porto, p.1-16, 30 out. 2006. Disponível em: <<http://ler.letras.up.pt/uploads/ficheiros/artigo5531.PDF>>. Acesso em: 30 nov. 2006.

SANTOS, Marcos Hercules ; NASCIMENTO, Lucileide Andrade de Lima do ; OLIVEIRA, Elias. Mineração em Meta-Dados Aplicado ao Processo de Desenvolvimento de Coleções. In.: SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 3., 2005, **Anais eletrônicos....** São Paulo: UNESP, 2005. Disponível em: <<http://bibliotecas-cruesp.usp.br/3sibd/docs/santos302.pdf>>. Acesso em: 20 ago. 2006.

SANTOS, Nilton Bahalis dos. **A ciência da informação e o paradigma holográfico**: a utopia de Vannevar Bush. 2005. 185 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

SENSO, José A.; PIÑERO, Antonio de La Rosa. El concepto de metadato. Algo más que descripción de recursos electrónicos. **Ci. Inf.**, Brasília, v. 32, n. 2, p.95-106, maio/ago. 2003.

SHADBOLT, Nigel; BERNERS-LEE, Tim; HALL, Wendy. The Semantic Web revisited. **IEEE Computer Society**, [S. l], p.96-101, maio/jun. 2006.

SILVA, Edna Lúcia de; MENEZES, Estera M. **Metodologia da pesquisa e elaboração de dissertação**. 3. ed. Florianópolis: Laboratório de Ensino à Distância da UFSC, 2001. 121 p.

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

TAKAHASHI, Tadao (org.). **Sociedade da informação no Brasil**: livro verde. Brasília: Ministério da Ciência e Tecnologia, 2000. 195 p.

VALE, Maria do Socorro C.; COSTA, Denise Coutinho; ALVES JR., Nilton. **Internet**: histórico, evolução e gestão. Disponível em: <<http://www.rederio.br/downloads/pdf/nt00501.pdf>>. Acesso em: 27 out 2006.

W3C: Leading the web to its full potential. [S.l.]: World Wide Web Consortium, [2005]. Disponível em: <<http://www.w3c.org/2005/05/At-a-Glance.pdf>> .Acesso em: 25 set. 2006. Panfleto.

WIKIPÉDIA. Desenvolvido pela Wikimedia Foundation. **Internet**. Disponível em: <<http://pt.wikipedia.org/w/index.php?title=Internet&oldid=3434743>>. Acesso em: 8 set 2006

WORLD Wide Web Consortium. Disponível em: <<http://www.w3.org/>>. Acesso em: 10 maio 2006.