UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Filosofia e Ciências,
Campus de Marília - SP

JOSÉ EDUARDO SANTAREM SEGUNDO

Representação Iterativa: um modelo para Repositórios Digitais

Marília – SP
2010

JOSÉ EDUARDO SANTAREM SEGUNDO

Representação Iterativa: um modelo para Repositórios Digitais

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP - campus de Marília, como requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Informação, Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Orientadora: Dra. Silvana Ap. Borsetti Gregorio Vidotti

Marília
2010

Santarem Segundo, José Eduardo
S233r Representação Iterativa: um modelo para Repositórios Digitais / José Eduardo Santarem Segundo. – Marília, 2010.
224 f. ; 30 cm.

Tese (Doutorado em Ciência da Informação). – Faculdade de Filosofia e Ciências , Universidade Estadual Paulista, 2010.
Bibliografia: f. 140-150
Orientadora: Vidotti, Silvana Aparecida Borsetti Gregório

1. Repositórios Digitas. 2. Representação Iterativa. 3. Folksonomia. 4. Folksonomia Assistida. 5. Web Semântica. 6. Recuperação da Informação. 7. Ontologia. I. Autor. II. Título.

CDD – 004.6

JOSÉ EDUARDO SANTAREM SEGUNDO

Representação Iterativa: um modelo para Repositórios Digitais

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP - campus de Marília, como requisito parcial para a obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Informação Tecnologia e Conhecimento.
Linha de Pesquisa: Informação e Tecnologia
Orientadora: Dra. Silvana Ap. Borsetti Gregorio Vidotti

Marília, 24 de fevereiro de 2010.

BANCA EXAMINADORA

_____
Profª Drª Silvana Aparecida Borsetti Gregorio Vidotti (Orientadora)
Universidade Estadual Paulista / UNESP

_____
Profª Drª Plácida Leopoldina Ventura Amorim da Costa Santos
Universidade Estadual Paulista / UNESP

_____
Prof. Dr. Ricardo César Gonçalves Sant'Ana
Universidade Estadual Paulista / UNESP

_____
Prof. Dr. Guilherme Ataíde Dias
Universidade Federal da Paraiba

_____
Prof. Dr. Marcos Luiz Mucheroni
Universidade de São Paulo / USP

Dedicatória

Dedico este trabalho a três pessoas especiais na minha vida:

A minha mulher Luciana, que me acompanha e me incentiva a cada dia, que luta, perde o sono, respeita as ausências e alegra as presenças, enfim, que me faz acreditar cada vez mais que o amor é possível e que só assim um homem se torna completo. Este trabalho tem muito do amor que ela sempre me oferece.

A minha filha Alícia, uma princesa doce e meiga, que nasceu junto com a ideia de enfrentar este desafio, que aprendeu a fazer seminários, escrever projetos e em alguns casos assistiu às disciplinas do programa. Com ela aprendi que o amor pode ser puro e verdadeiro.

Ao meu filho Raul, garoto de sorriso fácil e alegria contagiante, que chegou mais recentemente para acompanhar o último ano desta pesquisa, entretanto com tempo para também assistir algumas aulas do programa e ir cumprindo créditos.

Essas três pessoas me ofereceram toda a estrutura familiar de que sempre precisei, a eles recorri nos momentos de angústia, é com eles que encontro paz e alegria e é principalmente por eles que sempre busquei concluir com dignidade e alegria mais esta etapa da minha vida.

Luciana, Alícia e Raul: eu amo vocês.

Agradecimentos

Agradeço especialmente à Profa. Dra. Silvana Vidotti, por quem tive a honra de ser orientado, tanto no mestrado como no doutorado. Agradeço a confiança que ela sempre depositou em meu trabalho, a paciência com que tratou minha falta de tempo, a competência para conduzir as orientações de forma que fossem muito proveitosas, ao tempo dedicado a este trabalho, dadas as inúmeras atividades pelas quais é responsável. Enfim, por ter me aberto os olhos e me direcionado pelo caminho da pesquisa e da docência.

Agradeço imensamente à Universidade Estadual Paulista, instituição que, sem dúvida, tem sido a mais importante da minha vida, nos últimos anos. Foi trabalhando na Unesp que finalizei meus estudos de graduação e, na sequência, conclui a especialização. Foi a Unesp que me permitiu desenvolver e concluir o mestrado e agora o doutorado. Foi na Unesp que aprendi a ser profissional, a ter respeito pelo trabalho coletivo, a entender o funcionamento do ensino público, enfim, a Unesp me deu oportunidades que nunca havia imaginado ter em minha vida profissional. Se não bastasse, foi trabalhando na Unesp que conheci, me apaixonei e me casei com a mulher que me acompanhará para sempre e é mãe de meus dois filhos. Foi pela Unesp que tive oportunidade de iniciar minha carreira docente, à qual me dedicarei exclusivamente daqui em diante. Portanto, só tenho a agradecer pelos 13 anos em que este lugar foi minha segunda casa! Para que não fiquem dúvidas, meu MUITO OBRIGADO!

Aos meus pais, meus irmãos e outros familiares, que me incentivaram e souberam entender as ausências durante os últimos quatro anos.

Aos professores, Plácida Santos e Ricardo Sant'Ana, que muito contribuíram no processo de qualificação, além das ricas discussões e, claro, por todo o aprendizado, e a todos os professores do programa, que de certa forma contribuíram para o meu desenvolvimento e muito me ensinaram.

Aos colegas de trabalho, por entenderem a importância desta pesquisa e por colaborarem na realização das atividades, durante os períodos em que estive afastado para o desenvolvimento desta tese.

A professora Élide Feres pela revisão ortogrática.

A Caroline, pelo apoio, incentivo e pronto atendimento, sempre.

Aos colegas de turma, que estabeleceram ótimos debates durante a realização das disciplinas: Elvis, Liriane, Zeca, Cesar, Carlos, Luana, Walter, Aldinar, Lourdes, Mario, Fabiano, Rachel, Iuri, Miguel Maurício, e todos os outros que fizeram parte desta história.

A quem rege e permite tudo nesta vida.

“Nossa loucura é a mais sensata das
emoções; Tudo o que fazemos deixamos
como exemplos para os que sonham um
dia serem assim como nós:
loucos... mas felizes!”

Mário Quintana

SANTAREM SEGUNDO, J. E. **Representação Iterativa: um modelo para repositórios digitais**. 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

Resumo

A recuperação da informação tem sido muito discutida e abordada dentro da Ciência da Informação nos últimos anos, principalmente depois da explosão informacional gerada pela Internet. A busca por informação de qualidade e compatível com a necessidade do usuário tem sido tratada como obsessão, atualmente. A utilização da Internet indicou novos modelos de armazenamento de informações, como os repositórios digitais, que têm sido utilizados em ambientes acadêmicos e de pesquisa como principal forma de autoarquivar e, principalmente, de disseminar informação, porém com uma estrutura de informação que sugere melhor descrição dos recursos do que a própria Web e indica uma melhor recuperação da informação nestes ambientes. Os repositórios ainda não estão aptos a recuperar informação de forma semântica e contextualizada. Os novos paradigmas de Internet sugerem utilização dos recursos de Web 2.0 e também de Web 3.0, permitindo, respectivamente, interatividade e também estrutura de informação semântica. Desta forma o objetivo desta pesquisa é melhorar o processo de recuperação da informação, apresentando uma proposta de modelo estrutural no contexto da Web Semântica, abordando o uso de recursos da Web 2.0 e Web 3.0 em repositórios digitais, que permita recuperação semântica da informação, através da construção de uma camada de informação chamada Representação Iterativa. Através do modelo sugerido e proposto — Representação Iterativa — será possível adequar os repositórios digitais para que utilizem Folksonomia e também vocabulário controlado de domínio, de forma a gerar uma camada de informação iterativa, que possibilite retroalimentação da informação, além de recuperação semântica da informação, através do modelo estrutural desenhado para repositórios. O modelo sugerido resultou na efetivação da tese de que através da Representação Iterativa é possível estabelecer um processo de recuperação semântica da informação em repositórios digitais.

Palavras-chave: Repositórios Digitais, Representação Iterativa, Folksonomia, Folksonomia Assistida, Web Semantica, Recuperação da Informação, Ontologia.

SANTAREM SEGUNDO, J. E. **Representação Iterativa: um modelo para repositórios digitais**. 2010. 224 f. Thesis (PhD Degree in Information Science) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

Abstract

The information retrieval has been much discussed and addressed in information science in recent years, especially after the information explosion created by the Internet. The search for quality information and compatible with the need of user has been treated as an obsession now. The use of the Internet indicated a new type of store information, such as digital repositories, which have been used in academic and research as the main way to autoarchive, and especially to disseminate information, but with an information structure that suggests a better description resources than the Web itself and indicates a better retrieval of information in these environments. The repositories are not yet able to retrieve information in a semantic and context. The new paradigm suggests the use of Internet resources of Web 2.0 as well as Web 3.0, allowing, respectively, interactivity, and also the structure of semantic information. Thus the objective of this research is to improve the process of information retrieval, with a proposed structural model in the context of the Semantic Web, addressing the use of Web 2.0 and Web 3.0 in digital repositories, enabling semantic retrieval of information through construction of a layer of information called Representação Iterativa. The model suggested and proposed — Representação Iterativa — you can adapt to the digital repositories using Folksonomy and also controlled vocabulary of the field in order to generate an iterative layer information, which allows feedback information, and semantic retrieval of information, through the structural model designed for repositories. The model suggested resulted in the realization of the thesis that through Representação Iterativa is possible to establish a process of semantic retrieval of information in digital repositories.

Palavras-chave: Digital Repositories, Representação Iterativa, Folksonomy, Folksonomy Assisted, Semantic Web, Information Retrieval, Ontology.

Lista de Exemplos

EXEMPLO 1 – ALGORITMO DE BUSCA EM LARGURA 49
EXEMPLO 2 – ALGORITMO DE BUSCA EM PROFUNDIDADE 50
EXEMPLO 3 - MICROFORMATO HCARD 94
EXEMPLO 4 - MICROFORMATO HCALENDAR – REUNIÃO DO GRUPO DE PESQUISA 96
EXEMPLO 5 – SENTENÇA RDF 121
EXEMPLO 6 – ESTRUTURA DE ONTOLOGIAS 130
EXEMPLO 7 – TAG OWL:ONTOLOGY 132
EXEMPLO 8 – CLASSES OWL 133
EXEMPLO 9 – HIERARQUIA DE CLASSES 134
EXEMPLO 10 – CONSTRUÇÃO DE UMA CLASSE 134
EXEMPLO 11 – INDIVÍDUO 135
EXEMPLO 12 – OUTRO EXEMPLO DE INDIVÍDUO 135
EXEMPLO 13 – PROPRIEDADE DE OBJETOS 136
EXEMPLO 14 – PROPRIEDADE DE DADOS 137
EXEMPLO 15 – SUB-PROPRIEDADE OWL 137
EXEMPLO 16 – SUB-PROPRIEDADE DE DADOS APLICADA A INDIVÍDUO 137
EXEMPLO 17 – RESTRIÇÃO DE CARDINALIDADE 138
EXEMPLO 18 - CABEÇALHO EM OWL DA ONTOLOGIA OPENCYC 147

Lista de Figuras

FIGURA 1 - FÓRMULA DA SIMILARIDADE 32
FIGURA 2 - GRAFO SIMPLES E DESCONEXO 44
FIGURA 3 - LISTA DE ADJACÊNCIAS PARA GRAFO SIMPLES. 47
FIGURA 4 – MATRIZ DE ADJACÊNCIAS PARA GRAFO SIMPLES. 48
FIGURA 5 – TWITTER DO GOVERNADOR DO ESTADO DE SÃO PAULO – JOSÉ SERRA 61
FIGURA 6 - CANAIS RSS – TERRA 64
FIGURA 7 - TAG CLOUDS 68
FIGURA 8 - BUSCA DEL.ICIO.US 69
FIGURA 9 - DEL.ICIO.US 70
FIGURA 10 - ESTRUTURA DA WEB SEMÂNTICA (LAYERCAKE). 72
FIGURA 11 - VALIDAÇÃO WEB STANDARD DO SITE DA W3C BRASIL 87
FIGURA 12 - VALIDAÇÃO WEB STANDARD DOS PORTAIS UOL E UNESP 87
FIGURA 13 - SELO DE VALIDAÇÃO WEB STANDARD - PADRÃO XHTML 1.0, NO SITE DO W3C BRASIL. 88
FIGURA 14 - APLICAÇÃO DE WEB STANDARDS EM UM DOCUMENTO WEB. 89
FIGURA 15 - HCREATOR 95
FIGURA 16 - ADD-ON OPERATOR DO FIREFOX IDENTIFICANDO E DISPONIBILIZANDO INFORMAÇÕES SOBRE MICROFORMATO HCALENDAR 95
FIGURA 17 - AGENDA DO GOOGLE RECEBENDO E AGUARDANDO USUÁRIO SALVAR A INFORMAÇÃO DO MICROFORMATO DA REUNIÃO. 97
FIGURA 18 – GOOGLE MAPS (MAPA LOCALIZADO ATRAVÉS DO MICROFORMATO DO EXEMPLO 4). 97
FIGURA 19 – DUBLIN CORE VIEWER EXTENSION. 99
FIGURA 20 – DIAGRAMA RDF 120
FIGURA 21 – PROTÉGÉ 2000 143
FIGURA 22 – OPENCYC 148
FIGURA 23 - MODELO LÓGICO DE BANCO DE DADOS – DSPACE 157
FIGURA 24 - PARTE DO MODELO FÍSICO DO DSPACE. 158
FIGURA 25 - INSERÇÃO DE OUTRO PADRÃO DE METADADOS NA FERRAMENTA DSPACE. ÁREA ADMINISTRATIVA DO SOFTWARE. 161
FIGURA 26 - ALTERAÇÃO DO PADRÃO DC QUALIFICADO NA FERRAMENTA DSPACE. ÁREA ADMINISTRATIVA DO SOFTWARE. 162
FIGURA 27 – TABELAS COMMUNITY, COLLECTION E COMMUNITY2COLLECTION 171
FIGURA 28 – TABELA METADATAFIELDREGISTRY (DSPACE) 173
FIGURA 29 – TABELA METADATAVALUE – DSPACE 173
FIGURA 30 – BUSCA NO DEL.ICIO.US 179
FIGURA 31 – TABELAS PARA ARMAZENAMENTO DAS TAGS 184
FIGURA 32 – TABELA TAGS POPULADA 186
FIGURA 33 – TABELAS TAGS2TAGS E TAGS2ITEM POPULADAS 186
FIGURA 34 – REPRESENTAÇÃO ITERATIVA – VISÃO DETALHADA 188
FIGURA 35 – EXEMPLO DE PÁGINA DE RESULTADOS. 197
FIGURA 36 – NUVEM DE TAGS DO MICROBLOG TWITTER 200
FIGURA 37 – MATRIZ DE ADJACÊNCIAS E QUATRO ARTIGOS UTILIZADOS COMO EXEMPLO. 206
FIGURA 38 – REDE DE TAGS DE QUATRO ARTIGOS UTILIZADOS COMO EXEMPLO. 208

Sumário

1 INTRODUÇÃO 13

1.1 Definição do Problema de Pesquisa 16
1.2 Hipótese, Tese e Proposição da Pesquisa 17
1.3 Objetivos 19
1.4 Metodologia 20
1.5 Justificativa 20
1.6 Estrutura do Trabalho 21

2 RECUPERAÇÃO DA INFORMAÇÃO 24

2.1 O que é a recuperação da informação 25
2.2 O usuário e o sistema de recuperação 27
2.3 Modelos de recuperação da informação 28
2.3.1 Modelo Booleano 30
2.3.2 Modelo Vetorial 32
2.3.3 Modelo Probabilístico 34
2.3.4 Outros modelos de recuperação 35
2.4 Modelos Dinâmicos de Recuperação da Informação 35
2.4.1 Algoritmos Genéticos e Redes Neurais 36
2.5 Recuperação da Informação na Web 38
2.6 Grafos 43
2.6.1 Lista de Adjacências e Matriz de Adjacências. 45
2.6.2 Busca em profundidade e busca em largura. 48

3 FUNCIONALIDADES E RECURSOS TECNOLÓGICOS PARA WORLD WIDE WEB 53

3.1 Web 2.0: conceitos e funcionalidades 54
3.1.1 Interfaces Ricas 57
3.1.2 Inteligência Coletiva 58
3.1.3 Wikis e Blogs 59
3.1.4 Mashup 62
3.1.5 RSS (Really Simple Syndication) 63
3.1.6 Folksonomia 65
3.2 Web 3.0 – A Web Semântica 71
3.2.1 Metadados 77
3.2.2 Dublin Core 80
3.2.3 Web Standards 85
3.2.4 Microformatos 90

4 ONTOLOGIAS: CONCEITOS, LINGUAGENS E FERRAMENTAS 100

4.1 Definição de ontologia 101
4.2 Estruturas de Representação do Conhecimento 106
4.2.1 Vocabulário Controlado 106
4.2.2 Tesauro 109
4.2.3 Taxonomias 112
4.3 Composição e Construção de Ontologias 114
4.4 Linguagens de Marcação Semântica 117
4.4.1 RDF e RDF Schema 118
4.4.2 Simple HTML Ontology Extensions (SHOE) 123
4.4.3 Ontology Inference Layer (OIL) 123
4.4.4 DAML e DAML+OIL 125
4.4.5 Web Ontology Language (OWL) 127
4.4.5.1 Estrutura OWL – Namespaces 130
4.4.5.2 Estrutura OWL – Cabeçalhos 131

4.4.5.3 Elementos Básicos OWL – Classes 133
4.4.5.4 Elementos Básicos OWL – Indivíduos 134
4.4.5.5 Elementos Básicos OWL – Propriedades 135
4.4.5.6 Elementos Básicos OWL –Restrições em Propriedades 138
4.5 Ferramentas para desenvolvimento de ontologias 139
4.5.1 OilEd 140
4.5.2 OntoEdit 140
4.5.3 Chimaera 141
4.5.4 API Jena 142
4.5.5 Protégé 2000 142
4.5.6 Outras iniciativas 144
4.6 Construção Automática de Ontologias 144
4.7 Ontologias de Topo 146

5 REPOSITÓRIOS DIGITAIS DE INFORMAÇÃO CIENTÍFICA 149

5.1 A estrutura de informação dos repositórios digitais. 156
5.2 A recuperação de informação em repositórios digitais. 163
5.3 Os recursos e funcionalidades da Web 2.0 em repositórios digitais 164
5.4 Os recursos e funcionalidades da Web 3.0 em repositórios digitais 167

6 REPRESENTAÇÃO ITERATIVA, MODELO DE ESTRUTURA PARA DESCRIÇÃO, ARMAZENAMENTO, REPRESENTAÇÃO DE RECURSOS E RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS CIENTÍFICOS 169

6.1 Armazenamento – a relação entre Dublin Core e Banco de Dados 170
6.2 Folksonomia em repositórios digitais científicos 175
6.3 Representação Iterativa, estruturando o modelo 177
6.3.1 Folksonomia Assistida, enriquecendo a descrição do recurso 178
6.3.2 Armazenando as tags de forma estruturada 182
6.3.3 Iteratividade, a retroalimentação da informação 187

7 RECUPERAÇÃO DA INFORMAÇÃO NO MODELO DE REPRESENTAÇÃO ITERATIVA 194

7.1 Critérios para recuperação da informação na Representação Iterativa 195
7.2 Nuvem de tags 200
7.3 Rede de tags 204

8. CONCLUSÕES 209

8.1 Projetos Futuros 211

REFERÊNCIAS 213

1 INTRODUÇÃO

É inegável que o mundo tem passado por transformações nos últimos anos, principalmente as ocasionadas pelo uso das tecnologias. A chegada das Tecnologias da Informação e Comunicação (TIC) à casa das pessoas, sua mobilidade e meio de acesso a uma vida com muito mais informação têm transformado o pensar e o viver de grande parte da população.

São imensas as mudanças ocorridas nas últimas duas décadas, que fazem repensar conceitos e ações diariamente.

O Brasil tem acompanhado a mudança nas formas de acesso à informação.

Os números permitem verificar que a tecnologia está cada vez mais presente na casa do brasileiro. A relação de domicílios brasileiros que tinham computadores no final de 2005 e no final de 2008, conforme pesquisa do NIC.BR, confirma esse crescimento.

No ano de 2005, o número de casas equipadas com computador correspondia a aproximadamente 17% das residências brasileiras, comparados aos valores do ano de 2008, que apresenta 28% das residências brasileiras equipadas com pelo menos um computador.

Esses números revelam que o Brasil aumentou em mais de 60% o número de equipamentos em residências em apenas três anos.

Vive-se um momento em que a única constante é a certeza da mudança, e as inovações advindas com a Tecnologia da Informação e Comunicação têm papel preponderante neste cenário. Mas as principais mudanças não têm ocorrido em função de tecnologias específicas, mas da forma de se relacionar com elas, tanto como indivíduos, como grupos ou organizações (EVANS & WURSTER, 1999).

A Ciência da Informação tem participado efetivamente desta transformação, alavancada pelo uso das novas tecnologias da informação e, principalmente, da Internet.

A Internet é a tecnologia mais surpreendente das últimas décadas e através dela se tem construído um novo ambiente de informação e conhecimento, tornando-a objeto de muitos estudos e pesquisas, tanto da Ciência da Informação como de outras áreas do conhecimento.

Dentro do contexto da Ciência da Informação, a Internet tem atuado diretamente como elemento facilitador no processo de disseminação da informação e do conhecimento, incluindo o conhecimento científico, que deixou de estar disponível apenas nas revistas científicas e livros impressos e passou a utilizar a estrutura tecnológica da Internet para ser disseminado através das revistas eletrônicas digitais e dos repositórios digitais.

Conforme afirmam Castro e Santos (2008, p.2),

> A relevância tanto da Web quanto das bibliotecas digitais para os diversos ramos da ciência tem impulsionado pesquisadores e comunidades científicas a buscar soluções de integração, intercâmbio e entendimento semântico sobre os conteúdos que nelas circulam, a fim de proporcionar uma recuperação mais precisa, relevante e significativa para o usuário final.

Ainda, para os autores

> As bibliotecas digitais se caracterizam como ambientes facilitadores de acesso às informações, sem a limitação de espaço e tempo, uma vez que nessas o tratamento dado ao recurso informacional requer uma descrição de forma e de conteúdo legível por máquinas com resultados compreensíveis aos humanos. Desse modo, destaca-se a necessidade de um tratamento de forma e conteúdo adequado para a representação e para a apresentação de informações, visando uma recuperação mais eficiente. (CASTRO; SANTOS, 2008, p.2)

Estende-se a afirmação anterior aos repositórios digitais, uma vez que se defende que as bibliotecas digitais apresentam algumas semelhanças, em sua estrutura, aos repositórios digitais, objeto de estudo desta pesquisa.

A busca por informações tem aumentado consideravelmente em ambientes acadêmicos brasileiros, especialmente de nível superior. Grande parte dos alunos têm acesso direto à rede Internet, ocasionando uma constante troca de informações e de conhecimento.

O uso da Internet pelos cidadãos brasileiros também tem crescido consideravelmente nos últimos anos. Ao final de 2008, o índice de pessoas que acessaram a Internet foi de 43% da população total, e, ao analisar apenas os usuários com nível médio ou superior de instrução, esse número sobe para 63% e 89%, respectivamente (NIC.BR, 2008).

Além de a Internet estimular o acesso à informação, o cenário atual, baseado no desenvolvimento das tecnologias que englobam as funcionalidades denominadas Web 2.0, tem intensificado a relação usuário *versus* Internet, visto que esta permite a interatividade entre ambos e estimula o uso constante da rede.

Para Blattmann e Silva (2007, p.198),

> a Web pode ser considerada uma nova concepção, pois passou a ser descentralizada, na qual o sujeito tornou-se um ser ativo e participante sobre a criação, seleção e troca de conteúdo postado em um determinado site por meio de plataformas abertas.

Os conceitos a respeito das funcionalidades da Web 2.0 já estão sedimentados e têm sido amplamente utilizados na estrutura de construção dos sites, favorecendo o uso colaborativo e tornando a Web uma verdadeira plataforma para publicação e consumo de informação.

Esse novo formato interativo adotado pela Internet passou a fazer parte da vida dos usuários, como aconteceu com a própria Internet algum tempo atrás. Os conceitos e itens que dão sustentação às funcionalidades da Web 2.0 foram incorporados aos negócios. Assim, ferramentas como wikis e blogs já passaram a fazer parte do contexto de trabalho da grande maioria das empresas.

Além das funcionalidades que buscam uma inteligência coletiva e um novo patamar de interação, os conceitos também foram se

transformando em realidade no que diz respeito ao uso e a aplicação da Web Semântica.

Os conceitos da Web Semântica, cunhada por Tim Berners-Lee e homologada pelo W3C, têm sido objeto de estudo das Ciências da Informação e da Computação e despertado interesse da comunidade, de um modo geral. A Web 3.0, como tem sido chamada a Web Semântica, consiste num conjunto de padrões destinados a fazer com que o material publicado na Web possa ser recuperado de forma semântica, agrupando informações com o mesmo significado, independente de sua estrutura sintática, e permitindo associação de termos que são facilmente relacionados na estrutura cerebral do ser humano, porém são de difícil relacionamento em sistemas de informação.

Berners-Lee (2001) indica 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. (tradução nossa)

Através de recursos tecnológicos, a Web 3.0 tem efetivado uma mudança de paradigma em relação ao armazenamento e à recuperação de informações na Web.

1.1 Definição do Problema de Pesquisa

Diante de uma sociedade que interage de forma significativa com as novas funcionalidades representadas através das siglas Web 2.0 e Web 3.0, os ambientes informacionais digitais — bibliotecas e repositórios — de modo geral não incorporam tais tecnologias, que pode minimizar o interesse e, principalmente, o desenvolvimento de tais ambientes.

Se a Web, de modo geral, tem sido envolvida pela nova estrutura de informação, baseada nos conceitos da Web 2.0 assim como da Web 3.0, os usuários que estão sendo conduzidos à utilização de bibliotecas digitais e repositórios institucionais também passaram a ter o desejo de ver as tecnologias que compõem essas tecnologias aplicadas nestes ambientes.

As ferramentas utilizadas para instanciar bibliotecas digitais e repositórios digitais de informações apresentam em sua grande maioria uma estrutura que favorece, ou ao mínimo indica, o uso das tecnologias de Web 2.0 e Web 3.0, porém, em geral, não implementam esses recursos para que os usuários possam desfrutar desses benefícios em ambientes fechados e estruturados.

Visto que esse tipo de ambiente sugere criação de inteligência coletiva e tem como principal objetivo a disseminação da informação ciêntífica, questiona-se se a inserção das funcionalidades que compõem e nomeiam as tecnologias Web 2.0 e Web 3.0 não poderia contribuir com um avanço significativo no uso dos repositórios como plataforma universal no sentido de disseminar informação.

Cabe questionar também se é possível criar um ambiente que possa mesclar o uso das funcionalidades sugeridas nas tecnologias Web 2.0 e Web 3.0, visto que o primeiro conceito determina construção de inteligência coletiva de forma livre e a segunda sugere uso de um conjunto de termos de forma controlada, empregando uma ontologia de domínio que possa colaborar no sentido de caracterizar a recuperação semântica da informação.

Portanto, eis a questão principal: como incorporar os recursos e técnicas advindos das funcionalidades existentes nos conceitos de Web 2.0 e Web 3.0 em ambientes informacionais digitais como os repositórios digitais.

1.2 Hipótese, Tese e Proposição da Pesquisa

Baseado neste contexto, pode-se definir a tese levantada para esta pesquisa: a recuperação da informação em repositórios digitais no contexto da Web Semântica pode ser viabilizada por um modelo estrutural baseado na implementação de recursos da Web 2.0 e Web 3.0.

A hipótese desta pesquisa traduz-se na possibilidade de incorporar aos repositórios digitais uma arquitetura que permita o uso de Folksonomia Assistida, para autoarquivamento de objetos digitais, de forma que haja uma integração dos conceitos de Web 2.0 e Web 3.0, construindo um novo conceito de representação da informação — a Representação Iterativa —, de modo que possa propiciar aos usuários de ambientes reservados, como os repositórios digitais, armazenamento, descrição e, consequentemente, uma forma de recuperação mais contextualizada, com caráter dinâmico e semântico.

A Representação Iterativa é baseada na construção de uma camada de informação construída de forma social e cíclica, em que a estrutura ontológica vai sendo construída, inicialmente a partir de um conjunto controlado de termos, porém sujeita à reciclagem, de acordo com a ambiência e o conhecimento dos usuários utilizadores do sistema.

Para tanto, a proposta desta pesquisa é estabelecer um modelo de estrutura para repositórios digitais, que aplique conceitos de Web 2.0 e de Web 3.0. O modelo será baseado, principalmente, no uso de Folksonomia, que representa o uso de palavras-chave em formato aberto, mescladas com o uso de estruturas de representação do conhecimento, sistematizados e tecnologicamente utilizados no formato de ontologias, de forma que o sistema interaja diretamente com o usuário no momento da descrição do recurso, criando um conceito de Folksonomia Assistida em repositórios digitais de publicação, tornando esse repositório apto a proporcionar recuperação semântica de informações e a descrever os recursos informacionais de forma colaborativa, sugerindo um ambiente de construção coletiva de inteligência a respeito de um domínio de conhecimento e

construindo um novo modelo de estrutura informacional, baseado, principalmente, na experiência trazida através da interação do usuário.

1.3 Objetivos

Com intuito de contribuir com a área de Ciência da Informação, principalmente no que diz respeito aos ambientes informacionais digitais, essa pesquisa tem como objetivo melhorar o processo de recuperação da informação, apresentando uma proposta de modelo estrutural no contexto da Web Semântica, abordando o uso de recursos da Web 2.0 e Web 3.0 em repositórios digitais, que permita recuperação semântica da informação, através da construção de uma camada de informação chamada Representação Iterativa.

Dentro deste contexto, é possível dividir o objetivo geral em partes distintas que podem ser relacionadas da seguinte forma:

- Estabelecer uma modelagem complementar de banco de dados que possa caracterizar o uso dos conceitos de Folksonomia em repositórios digitais;
- Aplicar uma metodologia de sugestão a descrição de tags, sugerindo a utilização de uma representação amparada em Folksonomia e Estruturas de Representação do Conhecimento, que se denomina Folksonomia Assistida;
- Construir um corpus de informaçao baseado em uma ontologia de domínio e ampliada e reciclada com a experiência do usuário através do uso da Folksonomia Assistida, criando uma uma estrutura nomeada Representação Iterativa;
- Utilizar o modelo construído, Representação Iterativa, no processo de recuperação da informação em repositórios digitais, através da elaboração de uma rede baseada na teoria dos

grafos, de forma que seja possível recuperar informações com carater semântico.

1.4 Metodologia

O presente estudo caracteriza-se como uma pesquisa descritiva e analítica, com base em análise documental, dividida em duas partes: a primeira, caracterizada pela observação direta não participativa de ferramentas que implementam repositórios digitais, assim como de repositórios digitais já instanciados, visualizando tanto seu uso e seu comportamento quanto a questão de recursos relacionados às funcionalidades das chamadas Web 2.0 e Web 3.0; e a segunda, com característica exploratória, focalizando a proposição do modelo "Representação Iterativa: um modelo para Repositórios Digitais", onde sugere um modelo inovador para repositórios, com a utilização de estruturas de representação do conhecimento e participação do usuário na construção de um vocabulário próprio de domínio.

1.5 Justificativa

O estudo justifica-se pela necessidade atual de gerar uma estrutura de armazenamento e representação com utilização de funcionalidades que favoreçam a construção de uma inteligência coletiva nestes ambientes e, principalmente, uma recuperação mais apropriada de informação em um ambiente informacional digital. O atendimento a tais necessidades cria um novo conceito de publicação, descrição e armazenamento, dentro do formato tecnológico dos repositórios digitais, e sugere que conceitos estudados e definidos na Ciência da Informação sejam efetivamente aplicados e utilizados.

1.6 Estrutura do Trabalho

Além do presente capítulo, esta tese contempla o seguinte formato:

Capítulo 2 – Recuperação da Informação - Faz uma abordagem sobre recuperação da informação, visto que a principal motivação para o desenvolvimento da Web 3.0 e, principalmente, dos repositórios institucionais é permitir que os usuários tenham acesso mais qualificado e mais condizente com sua expressão de busca, oferecendo-lhes informações úteis para a construção de novos conhecimentos. Este capítulo apresenta uma contextualização teórica a respeito da evolução da recuperação da informação e de seus principais métodos, além de uma introdução à teoria dos grafos, que permitirá a recuperação da informação em estruturas em formato de redes.

Capítulo 3 – Funcionalidades e recursos tecnológicos para World Wide Web – Faz uma apresentação dos principais conceitos e técnicas que fundamentam e são responsáveis pelas funcionalidades que caracterizam ambientes Web 2.0 e Web 3.0. A abordagem principal incide sobre os recursos individualmente utilizados e considerados pontos-chave na construção de um ambiente colaborativo (Web 2.0) e semântico (Web 3.0).

Capítulo 4 – Ontologias: conceitos, linguagens e ferramentas – Com relação à Web 3.0, dar-se-á ênfase ao desenvolvimento de Ontologias. Dada a abordagem que será feita neste trabalho, decidiu-se destinar um capítulo apenas a este conceito. É importante ressaltar que a Representação Iterativa considera o uso de qualquer tipo de estrutura de representação do conhecimento, porém a ferramenta mais indicada para este fim são as ontologias. Serão abordados os conceitos relativos a teorias, práticas e ferramentas para desenvolvimento de ontologias, que são fundamentais no desenvolvimento da Web 3.0. Neste capítulo também serão apresentadas informações sobre a linguagem OWL, considerada pelo World Wide Web Consortium (W3C) como a linguagem mais completa para implementação de ontologias.

Capítulo 5 – Repositórios digitais de informação científica - utilizados como objeto de estudo central desta pesquisa. Este capítulo é destinado a fazer uma apresentação dos repositórios digitais, que são ambientes destinados à publicação e autoarquivamento de informações. O tema inclui uma abordagem teórica e histórica sobre repositório e, em seguida, apresenta a relação dos repositórios com a recuperação da informação e com as técnicas de Web 2.0 e Web 3.0, através de uma metodologia de observação direta não participativa. Este capítulo objetiva ainda encaminhar o trabalho para a construção do modelo sugerido na proposição, com aplicação das técnicas e conceitos de Web 2.0 e Web 3.0 de forma efetiva, a fim de proporcionar aos repositórios um ambiente rico e interativo para os usuários que publicam e consomem informações neste tipo de ambiente informacional.

Capítulo 6 – Representação Iterativa, modelo de estrutura para descrição, armazenamento, representação de recursos e recuperação da informação em repositórios digitais científicos – Apresentar-se-ão a metodologia assim como o conjunto de teorias e técnicas que serão propostas, visando um novo modelo de armazenamento e representação de informação, baseado numa nova estrutura funcional para os repositórios, aplicando os conceitos de sugestão de tags, pelo próprio sistema. Será demonstrada a aplicação real do modelo sugerido – Representação Iterativa - para repositórios, aplicando os conceitos de Folksonomia Assistida, com o objetivo de orientar o usuário no momento de introduzir informações acerca da descrição do recurso a ser publicado nos repositórios digitais.

Capítulo 7 – Recuperação da informação no modelo de Representação Iterativa - Tem como propósito apresentar um modelo de recuperação da informação, de forma a utilizar os conceitos de Web 3.0 e do modelo de representação iterativa sugerida e abordada no capítulo anterior, permitindo aos usuários do repositório digital uma recuperação da informação de forma semântica e contextualizada. Dessa forma, apresenta uma seqüência critérios para que o modelo desenvolvido possa apresentar de maneira mais precisa os resultados solicitados pelos usuários em seu termo

de busca. Neste capítulo são ainda apresentados mais duas aplicações como forma de apresentação de resultados aos usuários: a nuvem de tags e a rede de tags.

A seguir, no capítulo 8, constarão as conclusões, seguidas das referências.

2 RECUPERAÇÃO DA INFORMAÇÃO

Este capítulo apresenta um levantamento bibliográfico sobre recuperação da informação, desde a criação do termo e do conceito, passando pela relação dos usuários com os sistemas de recuperação. Apresenta ainda os modelos mais conhecidos de recuperação da informação e faz uma abordagem sobre a recuperação da informação na Web, como ponto-chave desta pesquisa.

É fato que, nos últimos anos, a recuperação da informação tem assumido um papel diferenciado dentro dos estudos da Ciência da Informação. Inserida no contexto do uso da informação, no objeto de estudo da Ciência da Informação, a recuperação aparece como elo final na busca pela apresentação ao usuário da informação mais adequada no menor tempo possível, modificando os fazeres da Ciência da Informação, a fim de proporcionar uma recuperação da informação mais adequada ao contexto e à necessidade do usuário.

E não é apenas no uso que a recuperação da informação está inserida, ela está indiretamente relacionada com representação, armazenamento, descrição, organização, preservação e acesso à informação. A representação e organização de itens de informação deveriam prover o uso, a preservação e o acesso a informação pelo interessado. Infelizmente, o acesso à informação necessária não é uma atividade simples.

Segundo Saracevic (1996, p.45),

> o trabalho com a recuperação da informação foi responsável pelo desenvolvimento de inúmeras aplicações bem sucedidas (produtos, sistemas, redes, serviços). Mas, também, foi o responsável por duas outras coisas: primeiro, pelo desenvolvimento da CI como um campo onde se interpenetram os componentes científicos e profissionais. Certamente, a recuperação da informação não foi a única responsável pelo desenvolvimento da CI, mas pode ser considerada como principal; ao longo do tempo, a CI ultrapassou a recuperação da informação, mas os problemas principais tiveram sua origem aí e ainda constituem seu núcleo. Segundo, a recuperação da informação influenciou a emergência, a forma e a evolução da indústria informacional. Novamente, a

> recuperação da informação não foi o único fator, mas o principal. Como a CI, a indústria da informação atualmente não é apenas recuperação da informação, mas esta é o seu componente mais importante.

Apesar de se vivenciar um momento diferente, com o apoio de novas tecnologias e amparados pelo uso da Internet, que vêm mudando a maneira de se pensar sobre a recuperação da informação desde o surgimento da Web, no início dos anos 90, a busca pelo melhor resultado na recuperação é algo que já vem sendo abordado na Ciência da Informação há algum tempo, dentro dos fazeres da Biblioteconomia.

Não obstante o direcionamento diferente, a teoria das cinco leis fundamentais da Biblioteconomia, definidas por Ranganathan, que sintetizadamente pode ser apresentada como o melhor livro no menor tempo, poderia ser assim adaptada: o resultado mais preciso, que atenda da melhor maneira o usuário, no menor tempo e com a maior quantidade de informações necessárias.

Como parte final de todo um processo de armazenamento, seguido do uso da informação armazenada, a recuperação da informação tem sido cada vez mais abordada na busca por sistemas de recuperação que atendam melhor a necessidade dos usuários em relação a qualidade do conteúdo em relação ao termo de busca. Desde a publicação do "Manual de Documentação", de Paul Otlet em 1937 (LÓPES YEPES, 1989) e do MEMEX de Vannevar Bush em 1945 (BARRETO, 2008), que diversos estudos vêm apresentando métodos e técnicas para evoluir o processo de recuperação da informação.

2.1 O que é a recuperação da informação

O termo "recuperação da informação" foi cunhado em 1951, por Calvin Mooers, quando criou o termo "Information Retrieval" e definiu os problemas a serem abordados por esta nova disciplina. A Recuperação de

Informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação. (MOOERS, 1951)

Com o passar do tempo, passou a ser muito mais comum verificar o termo recuperação da informação sendo tratado dentro de um modelo mais complexo denominado Sistemas de Recuperação da Informação (SRI). Esse modelo propõe todo o sistema de representação, armazenamento, gestão e recuperação da informação.

Para Lancaster e Warner (1993, p. 4-5), os SRIs são a interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários. Desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários. Lancaster (1968) já havia anteriormente registrado que os SRIs não informam o usuário no sentido de mudar seu conhecimento sobre objeto de sua questão, mas apenas o informam sobre a possível existência de documentos atinentes à questão, além de características desses documentos.

Baeza-Yates e Ribeiro-Neto (1999, p. 1) indicam que

> a recuperação da informação está diretamente ligada à representação, armazenamento, organização e acesso aos itens de informação. Dizem também que a representação e a organização dos itens de informação deveriam prover o uso e o fácil acesso a informação necessária ao usuário. (tradução nossa)

Portanto, desde 1951, com a primeira definição do termo por Mooers, a Recuperação da Informação vem sendo discutida, e novas técnicas e estudos desenvolvidos, a fim de buscar sempre o melhor resultado possível para o usuário que procura a informação.

A Ciência da Informação e a Ciência da Computação aparecem como as ciências mais envolvidas com a busca pela melhoria da qualidade da informação recuperada. A Ciência da Informação apresenta uma visão mais metodológica e tem procurado estruturar os dados e criar métodos e

modelos que proporcionem um melhor armazenamento da informação, assim como vem estudando metodos que agreguem semântica à informação, e consequentemente possam ser aplicadas no processo de recuperação. A Ciência da Computação tem procurado atuar na aplicação dos modelos citados, diretamente no desenvolvimento de técnicas computacionais, como algoritmos, que possam viabilizar as metodologias sugeridas e pesquisadas.

Apesar do envolvimento das duas ciências e de tantas pesquisas, o processo de recuperação ainda não conseguiu atingir a os resultados que os usuários precisam ou que os pesquisadores esperam e, portanto, continua abarcando pesquisadores ao redor do tema.

2.2 O usuário e o sistema de recuperação

A recuperação da informação pode ser vista por dois lados distintos que auxiliam o processo de busca da melhoria da informação recuperada. Baeza-Yates e Ribeiro-Neto (1999, p. 7) afirmam que o problema da recuperação da informação está entre duas visões, visão humana e visão computacional:

> [...] para a visão computacional o problema consiste principalmente na construção de índices eficientes, processamento de consultas (buscas) com alta performance, desenvolvimento de algoritmos que criem rankings e que recupere o melhor conjunto de resposta para a questão aplicada. A visão humana consiste principalmente no estudo do comportamento do usuário, na compreensão de suas principais necessidades e em determinar como a compreensão do usuário afeta a organização e operação dos sistemas de recuperação.

Assim se verifica que o processo de recuperar informação consiste não apenas em técnicas e métodos que envolvem desde o armazenamento até os algoritmos que providenciam a recuperação da informação, mas também em adaptar os sistemas baseado no comportamento do usuário nesse modelo de recuperação, entendendo como é a construção da

informação e, principalmente, como é a construção de suas instruções para recuperação da informação.

Nesse capítulo será abordada, principalmente, a visão computacional da recuperação da informação, e nas seções subsequentes, a recuperação da informação no contexto de aplicação em repositórios digitais, objeto de estudo desta pesquisa.

2.3 Modelos de recuperação da informação

A grande dificuldade no processo de recuperação da informação é conseguir atender à necessidade do usuário, indicando o que é mais ou então menos relevante dentro do contexto de sua consulta a um conjunto de informações. Apenas como ressalva, deve-se esclarecer que, em alguns casos, nem o próprio usuário sabe exatamente o que deseja encontrar.

Para exemplificar, sugere-se a desconsideração dos sistemas automatizados de recuperação da informação, e imagine-se uma grande caixa repleta de livros.

A essa caixa de livros, submeta um usuário para verificar o que lhe interessaria, de forma que pudesse manusear e consultar os livros disponíveis, selecionando os títulos que fossem importantes para sua pesquisa ou determinado trabalho.

Esse usuário teria dúvidas na escolha e, com certeza, poderia selecionar títulos que, posteriormente, talvez não atendessem a sua expectativa no contexto de sua necessidade de informação.

Esse pequeno exemplo mostra que a recuperação da informação é contemplada por muitos aspectos que certamente dificultam o processo de recuperação.

Aproveitando ainda o exemplo, poder-se-ia imaginar esse primeiro usuário, que já teve acesso anteriormente à caixa de livros, auxiliando um

segundo usuário com as mesmas necessidades de informação. Neste caso, a escolha dos livros pelo segundo usuário seria facilitada, pois além de poder ter o contato com o material, também teria a discussão com o primeiro usuário que já havia passado pela mesma experiência. Portanto, a discussão dos dois a respeito do conteúdo, além da facilidade do contato com o material, certamente facilitaria a seleção dos livros. Mesmo com o apoio do primeiro usuário, ainda assim não seria o suficiente para se ter a certeza de que os livros selecionados pelo segundo usuário seriam as melhores opções para atender às necessidades de informação desejada por eles.

Vendo a recuperação da informação sob esse prisma, percebe-se que as composições de modelos de recuperação se tornam cada vez mais necessárias, e, principalmente, que os métodos utilizadas no momento do armazenamento da informação são ainda mais importantes, pois quanto mais claramente for representado um conteúdo, teoricamente mais fácil de recuperar ou de fazer parte de uma seleção esta informação estará.

Para executar a recuperação da informação baseada na busca de termos, foram desenvolvidos vários modelos de recuperação da informação.

Ferneda (2003, p.18) afirma:

> A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que o mesmo utiliza. Um modelo, por sua vez, influencia diretamente no modo de operação do sistema.

Os modelos de recuperação da informação são apresentados por vários autores, e a grande maioria deles apresenta um agrupamento ou divisão entre os modelos. Os chamados modelos clássicos de recuperação da informação são os que apresentam estratégia de busca para uma consulta. Normalmente nesses modelos é considerado que cada documento é representado por termos de indexação, ou seja, palavras-chave.

Os principais modelos clássicos apresentados são: Modelo Booleano, Modelo Vetorial e Modelo Probabilístico, porém cada um apresenta alternativas de extensão com o objetivo de evoluir em funcionalidade e desempenho.

Outro grupo de modelos de recuperação são os modelos dinâmicos de recuperação da informação, abordados por Ferneda (2003, p. 55) da seguinte maneira:

> Nesta ótica, os usuários interagem e interferem diretamente na representação dos documentos do corpus, permitindo uma evolução ou uma adaptação dos documentos aos interesses dos usuários do sistema, percebidos através de suas buscas e da atribuição de relevância (e não relevância) aos documentos recuperados (relevance feedback).

Os modelos clássicos ainda são muito aplicados nos sistemas de recuperação e, por isso, alguns serão apresentados a seguir.

2.3.1 Modelo Booleano

A álgebra da comutação foi primeiramente estudada em detalhes por George Boole, daí o nome álgebra booleana. O modelo booleano é baseado na álgebra booleana e na teoria de conjuntos. Na Álgebra Booleana, cada documento é representado por um conjunto de termos de índice e dessa forma o índice aponta qual documento é mais relevante, indicando assim uma relevância de maneira ordenada (CARDOSO, 2004).

No modelo booleano, a recuperação é sempre baseada na coincidência entre os termos que fazem parte do índice do documento e os termos estabelecidos na consulta através de uma expressão lógica.

A relevância estabelecida na expressão lógica é obtida com a aplicação de operadores lógicos (E, OU e NÃO), mais usados na forma de língua inglesa como AND, OR e NOT. É possível criar consultas mais restritivas e, em alguns casos, mais ricas, dependendo da combinação utilizada nos termos. O resultado da busca é influenciado diretamente pela ordem seqüencial de execução das operações lógicas, portanto é muito importante que a estrutura da expressão lógica seja bem clara e definida, utilizando-se os operadores supracitados, assim como o recurso dos parênteses que tem preferência de execução nas expressões.

O modelo booleano, assim como os outros, apresenta algumas limitações que devem ser conhecidas:

- Sendo a recuperação baseada em similaridade e comparação binária, a utilização de duas expressões diferentes pode gerar resultados iguais sem diferenciação entre a relevância dos documentos recuperados. Em alguns casos, é nítido verificar que são recuperados, da mesma forma, documentos que têm similaridade em apenas um dos termos da consulta, assim como documentos que apresentam vários dos termos de consulta. O resultado não expressa a relevância entre esses documentos recuperados, tratando-os simplesmente como documentos recuperados.
- Não é possível, através do modelo booleano, apresentar resultados parciais, a estrutura binária de funcionamento sempre apresenta resultados exatos, baseados nas comparações binárias de 1 ou 0.

Principalmente pelas limitações do modelo booleano, a eficácia dos sistemas de recuperação nele baseadas geram desconfiança nos resultados apresentados, e estes passam a ser utilizados em parte ou em conjunto com outros modelos de recuperação.

As limitações apresentadas demonstram de maneira ainda mais clara que é necessário conhecer o modelo para formular uma boa expressão de busca, e que, portanto, quanto mais simples for a expressão, mais "binário" será o resultado.

Ferneda (2003, p. 24) afirma:

> Expressões complexas exigem um conhecimento profundo da lógica booleana e evidenciam a importância da elaboração de uma estratégia de busca adequada para garantir a qualidade da informação recuperada. O conhecimento da lógica booleana é importante também para entender e avaliar os resultados obtidos em uma busca.

A relação forte com conceitos vindos da matemática como ciência e a clara forma de apresentação estimulam ainda o uso dos operadores booleanos, porém não são suficientes para manter seu uso isoladamente.

2.3.2 Modelo Vetorial

O desenvolvimento do modelo vetorial, no ano de 1968, por Gerard Salton, foi motivado principalmente pelas limitações apresentadas no modelo booleano (SALTON, 1988).

Esse modelo tem como premissa considerar a similaridade parcial entre os termos, representando-os através de um vetor numérico, onde cada elemento do vetor representa um termo de consulta e a este é atribuído um peso que indica tamanho e direção do vetor de representação. São esses pesos que possibilitam a proximidade de consulta e o cálculo da similaridade parcial entre os termos da consulta e os documentos, possibilitando que os resultados sejam apresentados de maneira classificada, de acordo com o grau de similaridade entre o termo na expressão de busca e o documento recuperado. O cálculo de proximidade entre os vetores é realizado de acordo com o ângulo do vetor, e dessa forma é calculado o grau de similaridade de acordo com a seguinte fórmula:

$$sim(x,y) = \frac{\sum_{i=1}^{t}(w_{i,x} \times w_{i,y})}{\sqrt{\sum_{i=1}^{t}(w_{i,x})^2} \times \sqrt{\sum_{i=1}^{t}(w_{i,y})^2}}$$

Figura 1 - Fórmula da Similaridade
Fonte: Ferneda (2003, p. 30)

Onde:

- ***x*** e ***y*** são os vetores;

- ***t*** *é o número total de documentos considerados;*
- $w_{i,x}$ é o peso do i-ésimo elemento do vetor x;
- $w_{i,y}$ é o peso do i-ésimo elemento do vetor y;
- ***sim*** *é a função de similaridade.*

No modelo vetorial, a consulta é realizada em busca dos termos designados, e a classificação apresentada como resultado baseia-se na frequência dos termos no documento em relação ao peso atribuído a cada termo, utilizando-se o grau de similaridade calculado.

É importante ressaltar que a construção do vetor de termos deve ser a mais significativa possível e de preferência utilizar uma quantidade restrita de termos, facilitando a eficácia do modelo vetorial.

Segundo Salton e Buckley (1988),

> quando um conjunto grande de termos é utilizado para a representação de um documento é alta a chance desse documento ser considerado semelhante a outro documento ou consulta.

Desta forma, é importante que a quantidade de termos não interfira diretamente na qualidade da recuperação da informação.

O uso de similaridade e do modelo vetorial facilita diretamente o processo de definição de um ranking para os resultados da consulta. Souza (2006, p. 167) compartilha desta ideia da seguinte forma:

> O modelo vetorial é a base da grande maioria de sistemas de recuperação de informações, mais notadamente os que têm como objeto a Internet, embora estes utilizem também outras técnicas para determinar o ranking de documentos como resposta a uma consulta.

O modelo vetorial apresenta como principal característica a simplicidade e a facilidade com que permite calcular a similaridade entre informações genéricas, além de executar comparações parciais, diferente do modelo booleano, que aplica comparações exatas que permitem a criação de uma classificação ordenada (FERNEDA, 2003).

Esse modelo também apresenta restrições, entre elas destaca-se o fato de não permitir o uso da álgebra booleana dentro de seu contexto; além disso, caracteriza-se por aproximar muito as combinações, podendo encontrar relação entre termos que não têm nada em comum.

2.3.3 Modelo Probabilístico

A teoria das probabilidades teve início com os jogos de dados, cartas e roleta. Esse é o motivo da grande existência de exemplos de jogos de azar no estudo da probabilidade.

A teoria da probabilidade calcula a chance de ocorrência de um número em um experimento aleatório que, quando repetido em condições iguais, pode fornecer resultados diferentes, ou seja, são resultados gerados ao acaso. Os experimentos aleatórios podem ser representados por sorteios de loteria ou até por um simples lançamento de um dado (SALTON e BUCKLEY, 1988).

O modelo probabilístico foi proposto por Maron e Kuhns, em 1960. Esse modelo prevê a classificação de documentos de acordo com sua probabilidade, em relação aos termos aplicados na busca. Nele se verifica a relação de relevância da expressão de busca em relação a cada documento, para investigar a probabilidade de relevância entre eles, supondo que exista um conjunto ideal de documentos que atende a cada uma das consultas aos dados, e que esse conjunto pode ser recuperado.

Uma busca inicial em um conjunto de documentos e o retorno do usuário em cada uma das interações permite o refinamento contínuo em direção a melhores resultados, portanto o feedback do usuário é determinante para que nas próximas buscas o sistema possa aproveitar os resultados anteriores para considerar documentos relevantes nas consultas, ficando explícita a importância do usuário na recuperação da informação, utilizando o modelo probabilístico.

Salton e Buckley (1988, p.2) indicam que "[...] em 1977, Robertson analisou o modelo probabilístico e observou que um documento deveria ser recuperado se sua probabilidade de ser relevante for maior que a sua probabilidade de não ser relevante" (tradução nossa). Partindo do princípio da recuperação da informação, esse modelo recebeu o nome de *Binary Indepence Retrieval.*

O modelo probabilístico caracteriza-se, principalmente, por apresentar um bom desempenho quando aplicado, visto que as estimativas de probabilidade já apresentam resultados de classificação, que podem ser utilizadas para apresentação dos resultados; entretanto, é notável que o fato de não explorar a frequência dos termos é visto como ponto negativo do modelo.

2.3.4 Outros modelos de recuperação

Além dos já citados, alguns outros modelos alternativos foram criados com o intuito de melhorar a performance ou a qualidade de recuperação dos modelos matemáticos já descritos.

O modelo booleano estendido é um modelo alternativo aos modelos booleano e ao vetorial, visto que tem como premissa aplicar o modelo booleano dentro de um vetor de similaridade, aliando assim a flexibilidade do modelo vetorial e a precisão do modelo booleano.

Junto ao modelo probabilístico podem ser implementadas as redes de Inferência, que têm o papel de inserir, no modelo probabilístico, variáveis aleatórias ao processo de raciocínio, usando fontes de evidência que podem estabelecer relacionamentos entre consultas futuras e consultas já realizadas no conjunto de documentos.

2.4 Modelos Dinâmicos de Recuperação da Informação

Os modelos dinâmicos de recuperação da informação surgiram a partir de um fenômeno de esgotamento das funções e fórmulas matemáticas nos estudos sobre recuperação da informação.

Bentlet (2002) apresenta diversos modelos computacionais inspirados em processos biológicos, tais como as Redes Neurais e os Algoritmos Genéticos. Neste trabalho será apresentada uma introdução a estes métodos como forma de ilustrar o conceito de modelos dinâmicos de recuperação da informação.

2.4.1 Algoritmos Genéticos e Redes Neurais

Os algoritmos genéticos têm sido introduzidos na busca por melhores resultados na recuperação da informação.

Entres os motivos da escolha da apresentação deste método neste trabalho de pesquisa é o fato de que a utilização de algoritmo genético interage de forma significativa com o usuário, sendo que o comportamento do ser humano que está participando do processo de recuperação da informação é elemento importantíssimo nas buscas subsequentes.

O fato de esta pesquisa sugerir um modelo de representação de forma iterativa, ou seja, que se recicla através de uma participação do usuário, torna o algoritmo genético importante no processo de recuperação da informação e, principalmente, na confirmação da participação do usuário humano no algoritmo de recuperação da informação.

Ferneda (2009) afirma que:

> A aplicação dos conceitos de Algoritmos Genéticos permite o desenvolvimento de sistemas evolutivos, nos quais os usuários, através de suas buscas, são elementos efetivamente participantes do processo de representação dos documentos do corpus do sistema.

O algoritmo genético se baseia no fato de que todo novo ser é formado através de características herdadas de seu pai e da sua mãe, sendo

que este novo ser pode ter uma porcentagem maior ou menor de características de cada um de seus genitores.

Segundo Ferneda (2009),

> A cada iteração do algoritmo (*"geração"*), um novo conjunto de estruturas é criado através da troca de informações entre estruturas selecionadas da geração anterior. O resultado tende a ser um aumento da adaptação dos indivíduos ao meio ambiente, podendo acarretar também um aumento da aptidão de toda a população a cada nova geração, aproximando-se de uma solução ótima para o problema em questão.

A aplicação deste método na recuperação da informação sugere que o processo de recuperação pode ser aplicado de uma forma mais natural, tendendo a evoluir, deixando de aplicar apenas conceitos matemáticos que tenham como padrão a manutenção constante do método.

Ferneda (2009) afirma:

> A aplicação dos algoritmos genéticos em sistemas de informação representa uma nova forma de pensar o processo de recuperação de informação na qual as representações dos documentos são alteradas de acordo com a necessidade de informação da comunidade de usuários, manifestada através de suas buscas.

Portanto, dentro do contexto de informação que se tem presenciado na Web, a aplicação de algoritmos genéticos na recuperação de informação pode ser considerada uma promissora alternativa de busca.

As redes neurais, assim como os algoritmos genéticos, procuram melhorar o processo de recuperação através de interação com o ambiente em que estão inseridas. Essa característica de adaptação coloca-os na categoria de modelos dinâmicos, porque vão se adaptando com o passar do tempo.

Segundo Ferneda (2006, p.25),

> Redes neurais constituem um campo da ciência da computação ligado à inteligência artificial, buscando implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas. Nesse sentido, apresentam capacidade de adaptar os seus parâmetros como resultado da interação com o meio externo, melhorando gradativamente o seu desempenho na solução de um determinado problema.

O conceito principal de funcionamento do modelo de redes neurais está em procurar simular o processamento de informações utilizadas pelo cérebro. Elas são compostas por unidades que representam os neurônios do cérebro e que fazem ligações com outros neurônios através das chamadas conexões sinápticas.

Esse modelo pode ser representado por grafos ponderados, onde cada vértice pode representar um neurônio e as conexões sinápticas podem ser representadas pelas arestas, de forma que as ligações mais representativas podem ser pontuadas através da utilização de grafos ponderados.

As redes neurais artificiais se diferenciam pela sua arquitetura e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizado. A arquitetura de uma rede neural restringe o tipo de problema no qual a rede poderá ser utilizada, e é definida pelo número de camadas (camada única ou múltiplas camadas), pelo número de nós em cada camada, pelo tipo de conexão entre os nós (feedforward ou feedback) e por sua topologia (HAYKIN, 2001).

Dentro do contexto de modelos dinâmicos de informação, as Redes Neurais se caracterizam como uma metodologia interessante no processo de recuperação da informação, principalmente no processo de recuperação da informação na web, porque o sistema pode “aprender” com as características dos usuários e utilizar este aprendizado para oferecer-lhes um conjunto de informações que mais condizem com sua busca, baseado nos resultados que foram mais interessantes do que nas vezes anteriores em que se utilizou o sistema de busca.

2.5 Recuperação da Informação na Web.

Os modelos de recuperação vêm sendo apresentados há muito tempo como alternativa à busca de informação em um conjunto de

documentos. Porém, dentro de uma nova dimensão como a Internet, fica visível o esgotamento de alternativas com relação a esses modelos já conhecidos, visto que existe uma clara mudança do corpus de consulta. Com a introdução da Internet no contexto do usuário, passa-se a ter um depósito de informações muito mais amplo, que carrega consigo a ligação de documentos e informações através de links, criando uma interligação entre os documentos armazenados e disponíveis na rede.

Embora tenha sido projetada para possibilitar o fácil acesso, o intercâmbio e a recuperação de informações, a Internet foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica, e se apresenta como um imenso repositório de documentos que não atende devidamente quando se precisa recuperar aquilo de que se tem necessidade (SOUZA E ALVARENGA, 2004).

Baeza-Yates e Ribeiro-Neto (1999, p.8) definem a Web como uma imensa base de dados onipresente e desestruturada.

Diferente de outros suportes de armazenamento, a Internet apresenta um diferencial, pois não implica espaço físico, como nas bibliotecas e museus, para depósito do material a ser armazenado. A constante falta de tratamento da informação antes que ela seja depositada, gera um processo de depósito que proporcionará dificuldade de recuperação posterior.

Dentro deste novo paradigma, o gerador de conteúdo, que é o responsável por depositar informações na Internet, não tem a devida preocupação em tornar seu documento mais relevante para as pesquisas, quando no processo de armazenamento e descrição da informação. Portanto, o que poderia facilitar o processo de recuperação da informação se torna um dificultador, inibindo a agilidade e a confiabilidade nos sistemas de recuperação da informação.

Outro fator que dificulta o processo de recuperação de informações na Web é que grande parte das informações depositadas na rede está localizada em ambientes fechados, que não permitem acesso aos sistemas de

busca ou estão armazenadas em estruturas que não são alcançadas pelos sistemas de recuperação. Este último tipo de informação citada caracteriza-se por ser denominada Internet Invisível.

Não obstante, é perceptível a necessidade do usuário de realizar buscas cada vez mais precisas e, principalmente, estabelecer sistemas de recuperação de informação que sejam rápidos e confiáveis. Dentro deste contexto, houve uma clara aproximação das áreas de Ciência da Informação e Ciência da Computação.

Nos últimos anos, muitas pesquisas têm sido direcionadas para a recuperação da informação em ambiente Web, porém ainda é nítido que grande parte dos usuários da Internet tem como forma mais usual de busca e recuperação de informações as ferramentas disponibilizadas pelas empresas Google[1] e Yahoo[2]. As empresas citadas têm melhorado e contribuído com o desenvolvimento do processo de recuperação, pesquisando e aplicando novos métodos e conceitos que tendem a facilitar, agilizar e tornar eficaz o processo de recuperação da informação na Web.

Ainda assim, o funcionamento destas ferramentas baseia-se em disparar robôs de busca, a fim de construir um arquivo invertido de indexação e, posteriormente, completar a recuperação sintática, baseada na comparação sintática entre termos, com outros métodos que, aplicados ao resultado inicial nos índices, procuram classificar os resultados de forma mais precisa ao usuário.

Um arquivo invertido é composto de uma lista previamente classificada de palavras-chave, onde cada palavra-chave tem uma lista de apontadores para os documentos que contêm aquela palavra-chave.

A utilização de índices apresenta-se ainda como a forma mais viável de proceder consultas em dados, sem a necessidade de fazer uma consulta diretamente na web no momento da solicitação do usuário, o que seria irremediavelmente lento, ou ainda uma alternativa a ter os sites do

---

[1] http://www.google.com.br
[2] http://www.yahoo.com.br

mundo todo armazenado em um banco de dados muito grande, o que tornaria o processo extremamente caro e inviável.

De acordo com Ferneda (2003, p. 96),

> Em um acervo extremamente grande como é a Web é essencial uma indexação antecipada de seus documentos (páginas). A maioria dos mecanismos de busca da Web gera índices. Pelo caráter dinâmico da Web esses índices devem permanecer em constante processo de atualização.

Outras técnicas têm sido frequentemente aplicadas, buscando proporcionar melhores resultados. O método de PageRank é uma destas técnicas, e tem como princípio calcular a "importância" de um site de acordo com a quantidade e "importância" dos sites que apontam para ele. O PageRank, que há algum tempo foi proposto pelo Google, já é utilizado por outras ferramentas de busca e recebeu extensões que agregam peso, assim como no modelo vetorial de recuperação, para ponderar o conjunto de links que direcionam para o site ou página Web em questão.

Outra técnica empregada para melhorar o processo de recuperação é a utilização de perfis de usuários combinados com avaliação de utilização. Alguns sites têm sugerido ao usuário que se cadastre, oferecendo em contrapartida serviços extras, e dessa forma tornando possível a criação de um dossiê da utilização das ferramentas que estão disponíveis, além do tipo de informação que aquele usuário está acostumado a utilizar. De posse dessas informações, é possível aplicar um filtro e relacionar com as informações acessadas, para assim criar uma lista de resultado, baseada e classificada de acordo com o tipo de informação que o usuário está acostumado a usar ou a procurar.

Aliado ao processo de utilização de perfil de usuário, pode-se recorrer à associação da busca recente com termos e resultados que já tenham sido recuperados pelo mesmo usuário ou ainda outro usuário que tenha características iguais ou semelhantes. Neste caso, vai se formando uma inteligência de pesquisa baseada nas recuperações de informações

anteriores. Essa técnica só poderá ser aplicada quando for possível armazenar e avaliar o perfil do usuário.

Para o método descrito, de análise das buscas anteriores, é possível dizer que toda vez que se faz uma busca e se obtém um resultado, se o usuário clica em um dos links de resposta e em segundos retorna novamente para a página de resposta da busca, pode-se afirmar que o resultado apresentado não é pertinente para aquela expressão de busca. Se, por outro lado, o clique direcionar a um site e, consequentemente, o usuário demorar a voltar ao site com os links de resposta, a ferramenta de busca deverá considerar esse site como importante para aquela pesquisa e utilizar em outras pesquisas que empreguem o mesmo termo.

Outra novidade em relação à recuperação da informação para Internet está na mistura de formatos de dados disponíveis na rede. Se há pouco tempo a Internet era carregada de arquivos em formato texto, essa tendência tem mudado fortemente nos últimos anos, passando a ter um conteúdo muito mais heterogêneo. Atualmente, impulsionados por aplicações como YouTube[3] e Flickr[4], há um volume maior de conteúdo disponível na Internet em formato de vídeo, áudio e imagens, além das habituais páginas em formato textual.

Essa nova característica no formato do material armazenado também representa uma dificuldade a mais no processo de recuperação e impacta diretamente nos modelos de recuperação da informação, visto que eles privilegiam principalmente a comparação sintática textual.

Notadamente, ainda no contexto da Internet, percebe-se um claro aumento de ambientes que têm se caracterizado por procurar organizar de forma mais clara e significativa as informações depositadas. As bibliotecas e os repositórios digitais são exemplos desses ambientes. Essas ferramentas tecnológicas têm sido utilizadas muito mais frequentemente com o passar dos anos.

---

[3] http://www.youtube.com
[4] http://www.flickr.com

Alguns ambientes, como repositórios digitais, têm uma estrutura bem definida para armazenamento de informações na Web, o que tende a facilitar o processo de recuperação.

O indicativo de que a recuperação da informação em bases textuais torna-se mais fácil e precisa em ambientes estruturados deve estar aliado ao cuidado dedicado ao processo de armazenamento, quando o documento a ser inserido na base deve ser muito bem catalogado e o conjunto de informações que caracterizam o documento deve estar muito claro para o sujeito que estará realizando o processo de postagem do material.

2.6 Grafos

No desenvolvimento desta pesquisa, foi avaliado o uso de grafos para auxiliar no processo de estruturação e recuperação da informação.

Grafo é um modelo matemático muito usado nas mais variadas formas de resolução de problemas, sendo apresentado na forma de um diagrama composto por pontos/círculos e linhas que unem esses círculos. Aos pontos é dado o nome de vértice e as linhas são conhecidas como *edges* ou arestas.

Goodrich e Tamassia (2002, p. 490) assim descrevem os grafos:

> Visto de forma abstrata, um grafo G é simplesmente um conjunto V de vértices e uma coleção E de pares de vértices de V, chamados de arestas. Assim, um grafo é uma forma de representar conexões ou relações entre pares de objetos de algum conjunto V. Alguns livros usam uma terminologia diferente para grafos e referem-se ao que chamamos de vértices como nodos e o que chamamos de arestas como arcos.

A teoria dos grafos é aplicada de forma sistemática desde que foi inventada no século XVII. Os primeiros trabalhos em teoria dos grafos surgiram no século XVIII. Vários autores publicaram artigos neste período, com destaque para o problema descrito por Euler, conhecido como As Pontes de Konigsberg (FEOFILOFF, KOHAYAKAWA e WAKABAYASHI, 2009).

Quando uma aresta liga dois vértices, os vértices são considerados adjacentes.

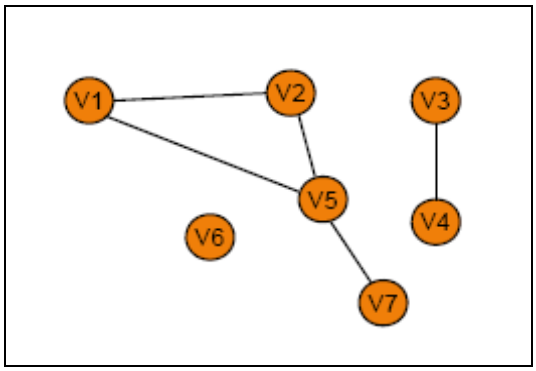


Figura 2 - Grafo simples e desconexo
Fonte: Próprio autor

A figura 2 apresenta um grafo simples. Quando um grafo possui mais de uma aresta interligando os mesmos dois vértices diz-se que este grafo possui arestas múltiplas (ou arestas paralelas), recebendo o nome de multigrafo ou grafo múltiplo. Um “grafo simples” não possui arestas múltiplas nem laços.

Matematicamente um grafo pode ser representado por G = (V,E), indicando que um grafo consiste de um conjunto de vértices (vertices) V, ligados por um conjunto de arestas (edges) ou arcos E. A figura 2 pode ser apresentada da seguinte forma:

- V(G) = {v1,v2,v3,v4,v5,v6,v7}
- E(G) = {(v1, v2); (v1,v5); (v2,v5); (v3,v4); (v5,v7)},

Onde:

- V(G), representa os vértices do grafo, apresentados dentro de um conjunto.
- E(G), representa as arestas, apresentadas através de pares ordenados entre os vértices, indicando que há ligação entre os vértices.

Os grafos podem ser conexos ou desconexos. Para que um grafo seja considerado conexo, todos os vértices devem ter ligação, mesmo que através de outro vértice, ou seja, é possível iniciar um caminho em um determinado vértice e chegar a qualquer outro. Qualquer grafo que tenha vértices, ou conjunto de vértices, em que não seja possível iniciar um caminho por eles e chegar a qualquer outro, é considerado desconexo.

Os grafos podem ser orientados ou não orientados. Grafos orientados são aqueles cujas arestas se apresentam com setas nas pontas, indicando a direção da aresta. Na figura 2 há um grafo não orientado, porque as arestas não têm direção, ou seja, não possuem setas. Na representação matemática das arestas de um grafo direcionado, os pares ordenados (i,j) e (j,i), onde i e j são vértices do grafo, são considerados diferentes.

Dependendo da necessidade do projeto ou do problema, os grafos também podem ser utilizados com pesos nas arestas, neste caso é atribuido o nome de grafo ponderado. No caso de grafos ponderados, os pesos são atribuídos às arestas, indicando uma maior ou menor densidade em relação à ligação entre os vértices ligados.

Esta tese propõe o uso de grafos, de forma que através deles será construída uma rede de elementos que será modificada a cada novo depósito de um documento no repositório digital e que auxiliará o usuário a realizar a recuperação da informação no modelo proposto.

Mesmo com as informações armazenadas em um banco de dados, para que se possa aplicar algoritmos de busca e recuperação da informação em grafos é necessário utilizar modelos computacionais como listas e matrizes de adjacências.

2.6.1 Lista de Adjacências e Matriz de Adjacências.

Para representar um grafo são necessários dois conjuntos: um para armazenar os vértices e o outro para armazenar as arestas. Estes dois conjuntos que formam um grafo podem ser representados por duas estruturas computacionais: lista de adjacências e matriz de adjacências.

Dois vértices são adjacentes quando existe uma aresta entre eles, portanto para vértices *i* e *j*, podemos dizer que temos um par ordenado e(i,j), que representa a adjacência.

A lista de adjacências é a forma de representação mais compacta para os grafos e sua construção se dá de forma que um grafo G usa um vetor com N listas ligadas, sendo que cada posição do vetor corresponde a um vértice do grafo, G(V,E), ficando as arestas representadas por listas ligadas.

Goodrich e Tamassia (2002, p. 502) confirmam o desempenho do uso de lista de adjacência:

> A lista de adjacência provê acesso direto tanto das arestas pra os vértices quanto dos vértices para suas arestas incidentes. Ser capaz de prover acesso entre vértices e arestas em ambas as direções permite-nos acelerar o desempenho de uma série de métodos para grafos se usarmos lista de adjacência.

A figura 3, no seu primeiro desenho, apresenta um grafo, e no segundo desenho, a representação em forma de lista de adjacências do grafo. Verifica-se que há um vetor como base na vertical, indicando que cada posição do vetor serve para representar um vértice do grafo. A partir de cada posição do vetor inicia-se uma lista ligada que serve para indicar quais são as adjacências do vértice em questão.

A representação matemática da figura 3 dá-se da seguinte forma:

- V(G) = {a,b,c,d,e }
- E(G) = {(a,b); (a,e); (b,c); (b,d); (b,e); (c,d); (d,e)},

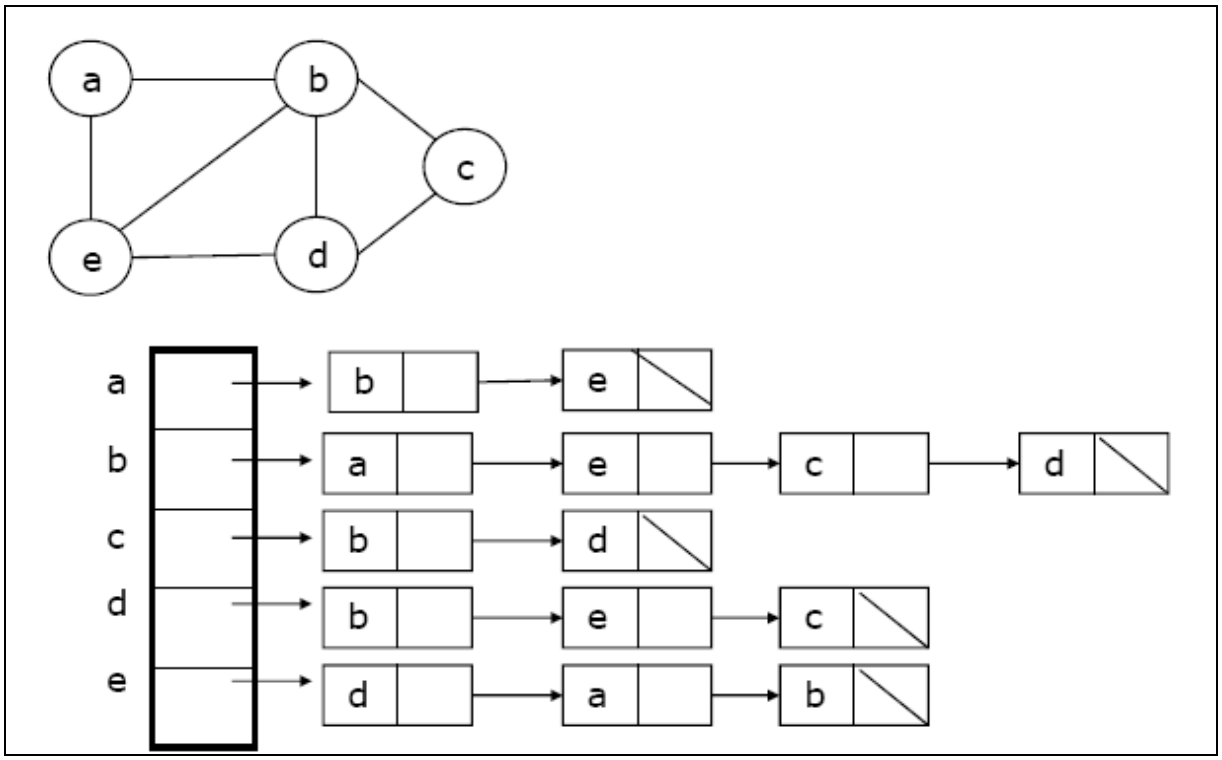


Figura 3 - Lista de adjacências para grafo simples.
Fonte: Próprio autor

No caso de grafos ponderados, poder-se-ia criar a lista ligada com dois campos, sendo que o segundo campo poderia carregar o peso/valor da aresta.

A matriz de adjacências é outra estrutura utilizada para armazenar informações de grafos. Para construir a matriz de adjacências para um grafo G=(V,E), assume-se que os vértices são identificados da seguinte forma: a, b, c, ..., Y, sendo Y o número total de vértices. Constrói-se uma matriz de adjacência com dimensão Y x Y e elementos e_{ij}, cujo valor pode ser 1 se (i,j) pertence a E e 0 se (i,j) não pertence a E, conforme pode ser observado na figura 4.

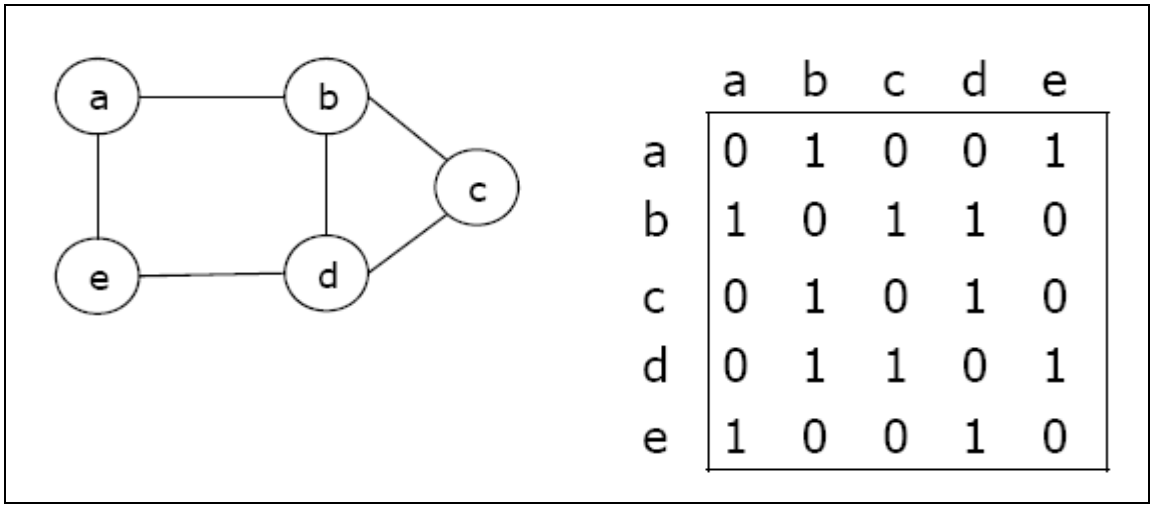


Figura 4 – Matriz de adjacências para grafo simples.
Fonte: Próprio autor

A indicação com o valor 1 para a representação de que existe uma aresta entre dois vértices pode ser alterada para um outro valor, representando o peso de uma aresta, no caso de grafos ponderados.

No caso de grafos orientados, é preciso observar o sentido do caminho entre os nós e adotar um padrão para o sinal dos pesos.

Nos grafos simples há uma simetria entre os elementos da matriz, portanto, com o objetivo de economizar memória, pode-se optar por armazenar apenas a matriz triangular inferior ou superior.

Através das estruturas apresentadas, é possível caminhar pelos grafos através de seus vértices e arestas, percorrendo caminhos em busca de informações.

Através de algoritmos, é possível determinar o procedimento para traçar um caminho dento de um grafo. Neste trabalho dar-se-á ênfase ao uso da busca em profundidade e da busca em largura como forma de explorar.

2.6.2 Busca em profundidade e busca em largura.

Os métodos de busca em largura e profundidade em grafos são formas sistemáticas para realizar a exploração dos vértices de um grafo, com o objetivo de se obter informações sobre a estrutura, ou seja, a busca é um método baseado em um algoritmo para caminhar pelos vértices e arestas de um grafo.

Um dos métodos utilizados é a busca em largura. A ideia principal de uma busca em largura consiste em processar todos os vértices de um determinado nível antes de ir para o próximo nível. Todos os nós localizados a uma distância **d** de um nó **n**, escolhido de forma aleatória, são percorridos antes dos nós localizados a uma distância **d+1** de **n**;

Goodrich e Tamassia apresentam aqui um algoritmo de busca em largura, com o uso de filas, que são estruturas de dados computacionais onde a primeira informação que entra na fila deve ser a primeira a ser retirada, como se fosse uma fila de banco.

```
Inicializa a lista L0 para conter o vértice s
i <- 0
enquanto L0 nao estiver vazia faça
   crie a lista Li+1, inicializando-a vazia
   para cada vértice v em L faça
      para cada aresta e incidente a v faça
         se aresta e for inexplorada entao
           seja w o outro ponto final de e
           se o vértice w é inexplorado entao
              rotule e como uma aresta de descoberta
              insira w em Li+1
           senao
              rotule e como uma aresta de cruzamento
      i <- i+1
```

EXEMPLO 1 – ALGORITMO DE BUSCA EM LARGURA

Conforme pode ser observado no algoritmo, a ideia da busca em largura é alcançar todos os vértices de um determinado nível para só depois passar para o próximo nível em busca de novos vértices, daí o nome de busca em largura.

O outro método utilizado para passeio nos grafos é a busca em profundidade, que consiste em sempre procurar “de forma vertical” novos vértices, até que seja possível atingir o ultimo nível.

Dessa forma, o procedimento para um nó **n**, escolhido de forma aleatória, visita-se um de seus nós adjacentes. E para cada um desses nós que for visitado, visita-se um dos nós adjacentes, e assim por diante, até o momento que for encontrado um nó sem adjacentes. Então, ocorre um “retorno” (backtracking) com o objetivo de visitar os nós restantes adjacentes a **n**, e o processo repete-se novamente.

Para o armazenamento de dados, a busca em profundidade utiliza uma estrutura computacional chamada pilha, onde a primeira informação armazenada será a ultima a ser retirada. Como exemplo de pilha, cita-se a própria pilha de pratos, sendo que o primeiro a ser colocado na pilha será o ultimo a ser retirado.

```
Visita-se um nó, selecionado de forma aleatória.
    Em seguida, o nó é marcado e empilhado em uma pilha s;
Enquanto a pilha s não estiver vazia:
    O nó n é desempilhado da pilha s;
    Para cada nó m (não marcado) que for adjacente a n:
        O nó m é visitado;
        O nó n colocado na pilha s;
        O nó m é colocado na pilha s;
        O nó m é marcado;
        Executa-se n ← m
```

EXEMPLO 2 – ALGORITMO DE BUSCA EM PROFUNDIDADE

Tanto a busca em largura quanto a busca em profundidade podem ser utilizadas na recuperação da informação, dependendo do processo desejado para percorrer os vértices de um grafo.

O novo perfil de usuário, os novos paradigmas de comunicação, a explosão informacional e as novas tecnologias da comunicação representam uma grande mudança em relação ao início dos anos 60, quando apareceram os primeiros catálogos online e quando ainda o poder de processamento das

máquinas era um tanto quanto limitado, mas cada um em sua época marcou de forma representativa a importância da tecnologia no processo de recuperação da informação. Recentemente, as pesquisas em Recuperação da Informação têm sido exploradas de forma mais significativa, em virtude da velocidade imposta pelo rápido desenvolvimento da Internet e a facilidade de acesso à rede, de um grande numero de usuários.

A forma diferente com que são tratados os ambientes Web em relação aos antigos ambientes que utilizavam os sistemas tradicionais de recuperação da informação traz ainda um novo contexto, que é a heterogeneidade no tipo de informação, além da globalização, que permite, cada vez mais, uma rede intrínseca de informações nos mais variados idiomas, e em alguns casos com o grande aproveitamento de palavras de um idioma por idiomas diferentes, criando um sério problema para os sistemas de recuperação que se baseiam em comparações sintáticas entre termos.

Mesmo em ambientes estruturados, como bibliotecas digitais e repositórios que tem um perfil organizacional muito mais adequado à recuperação da informação, a recuperação da informação ainda não satisfaz à necessidade do usuário, principalmente porque falta a esse processo uma técnica que permita fazer relações entre informações de forma semântica.

Neste capítulo, quando foram apresentados os modelos clássicos e dinâmicos de recuperação da informação, posteriormente sobre a disponibilização de informações na Internet e ainda uma pequena introdução à teoria dos grafos, ficou evidente a necessidade de estudos sobre comparações semânticas entre termos.

A globalização leva também a alguns novos problemas que não eram tão abordados, como polissemia e sinonímia.

No próximo capítulo serão apresentados os conceitos Web 2.0 e Web 3.0, criando uma nova estrutura de informação na Web. A abordagem sobre Web 3.0 ou Web Semântica apresenta o uso de ontologias como um dos caminhos para a formalização de uma estrutura que permita

comparação semântica e, principalmente, a organização e relação entre termos que sintaticamente não apresentariam nenhuma relação.

Esse novo conceito muda a forma de armazenar e descrever informações e altera a estrutura de recuperação da informação, afirmando a necessidade de constante revitalização dos sistemas de recuperação da informação.

3 FUNCIONALIDADES E RECURSOS TECNOLÓGICOS PARA WORLD WIDE WEB

Dentro do contexto evolutivo da Web, este capítulo tem como característica a apresentação das funcionalidades e conceitos relativos às tecnologias nomeadas Web 2.0 e Web 3.0, perfazendo uma apresentação dos componentes básicos e necessários que constituem esse novo modelo de Web.

No âmbito da Web 2.0 serão apresentadas as funcionalidades mais utilizadas, com destaque para a Folksonomia, que será abordada também em capítulos posteriores e na fase de implementação do modelo proposto nesta pesquisa.

No âmbito da Web 3.0 serão abordados os requisitos definidos por Tim Berners-Lee, pai da Web Semântica, para a construção desta. Em seguida, serão apresentados os Microformatos, como exemplo de aplicação de Web 3.0, mas que também podem ser considerados como aplicação de Web 2.0.

A evolução da Internet tem sido marcada nas últimas duas décadas por mudanças constantes de paradigma. Desde a criação do Mosaic — primeiro browser para navegação na Internet, desenvolvido por um grupo liderado por Marc Andreessen — até os dias de hoje, são constantes as mudanças e inovações.

Em princípio, foram as imagens e links que impulsionaram o mundo, construindo web sites que pudessem apresentar instituições, empresas e negócios, tornando a Web um dos recursos mais importantes para a divulgação de informações. Em seguida, vieram a evolução dos browsers e as novas linguagens para adicionar recursos à linguagem HTML, contribuindo, de certa forma, para o desenvolvimento da Web.

Com o passar do tempo, novas soluções surgiram, inclusive com o aparecimento da bolha “pontocom” da Internet, fenômeno observado entre 1995 e 2001, onde instituições, empresas e grupos dos mais variados ramos de atividade passaram a transformar o mundo virtual, canalizando

investimentos para sites, produtos e serviços na rede, dando início ao processo de e-commerce, com promessas de um caminho sem volta, cheio de possibilidades, que vem cada dia mais se consolidando.

A bolha "pontocom" levou as chamadas empresas de tecnologia a terem seus valores de mercado muito acima do que realmente valiam e podiam oferecer, inclusive com a criação de uma bolsa de valores específica para as empresas de tecnologia, a Nasdaq. O que parecia ser um caminho perfeito para algumas empresas teve fim com o estouro da bolha "pontocom", que culminou com a quebra de grande parte das empresas de tecnologia e a solidificação de empresas que já tinham uma boa estrutura de funcionamento.

O passar dos anos, os investimentos, as novas tecnologias e a massificação do uso da Internet como principal fonte de recursos de informação e de comunicação criaram a necessidade de mudanças, que vieram através dos novos conceitos apresentados através das funcionalidades da Web 2.0 e da Web Semântica, posteriormente chamada também de Web 3.0, que tem como princípios aproximar e facilitar o uso da Web pelos usuários.

3.1 Web 2.0: conceitos e funcionalidades

O termo Web 2.0 surgiu durante uma conferência[5] promovida pelas empresas de mídia Media-Live e O'Reilly Media, realizada em São Francisco, em outubro de 2004. Nesta conferência discutiu-se a ideia de que a Web deveria ser mais dinâmica e interativa, de modo que os internautas pudessem colaborar com seus conteúdos. Assim, começava a nascer a segunda geração de serviços online e o conceito da Web 2.0, onde surge um

nível de interação em que as pessoas podem colaborar para a qualidade do conteúdo disponível, produzindo, classificando e reformulando o que já está disponível.

Neste evento, em palestra de abertura, John Battle e O'Reilly fizeram uma lista preliminar de princípios em que o primeiro era "A Web como plataforma". Desde então, a idéia de Web 2.0 passou a ser discutida como sendo mais dinâmica e interativa, onde o foco não estava na tecnologia, mas na nova forma em que o usuário utiliza a Internet de modo colaborativo, com a criação de conteúdos.

Neste novo modelo, o usuário passa a ser o centro das atenções, ou seja, muda-se o paradigma e inicia-se uma nova concepção, que passa agora a ser descentralizada, de forma que o usuário se torna um participante ativo sobre a criação e seleção do conteúdo postado em um determinado site, através de plataformas abertas. Então, ao invés de apenas visualizar informações em páginas Web, os usuários podem publicar conteúdos em seus próprios blogs, em wikis e sites que compartilham fotos e vídeos. Os usuários passam a estabelecer colaboração ativa na rede, inserindo e combinando dados, conteúdos e serviços de várias fontes, para criar experiências e aplicativos personalizados.

O cerne da Web 2.0 está na intensa participação do usuário e na sua interatividade com os serviços on-line, muito mais voltada para a coletividade do que propriamente para o tecnológico, transformando a Internet em um espaço democrático de expressão e de acesso a todos, permitindo a construção da informação de maneira coletiva.

Dessa forma, o que efetivamente caracteriza a Web 2.0 é a participação ativa de usuários para: publicação, compartilhamento, organização e interação na construção da informação.

De acordo com Primo (2006, p.2):

> A Web 2.0 tem repercussões sociais importantes, que potencializam processos de trabalho coletivo, de troca afetiva,

[5] http://web2con.com

> de produção e circulação de informações, de construção social de conhecimento apoiada pela informática. São essas formas interativas, mais do que os conteúdos produzidos ou as especificações tecnológicas em jogo[...]

Neste novo contexto, tudo o que for realizado pelos usuários fica disponível na Web e pode ser acessado a qualquer momento por outros usuários ao redor do mundo, sem a necessidade de gravar em um determinado computador os registros de uma produção ou alteração na estrutura de um texto. As alterações são realizadas automaticamente na própria web.

De acordo com O'Reilly (2005, p.1),

> não há como delimitar fronteiras para a Web 2.0, pois trata-se de princípios e práticas para que diversos sites sigam. Um dos princípios fundamentais é a web como plataforma, ou seja, o usuário poder realizar atividades online que antes só eram possíveis com programas rodando em seu computador. O autor enfatiza que além da melhora na usabilidade e participação, o sistema também é incorporado por interconexão e compartilhamento.

Vários são os exemplos de utilização dos conceitos de Web 2.0. Rapidamente, podem-se citar ferramentas de conhecimento geral que iniciaram o processo de apresentação destes conceitos e outras que posteriormente aderiram à fórmula.

Os serviços da Google, como Orkut[6], Gmail[7], Blogger[8], utilizam tecnologias, como Ajax, Javascript, XML, além de outros, como Del.icio.us[9], um gerenciador de bookmark, o Flickr, que, além de permitirem a hospedagem de fotos, também possibilitam organizá-las através de associações livres, registrando as fotos conforme o título que o depositante interprete como sendo o mais adequado.

A Web 2.0 apresenta, como se pode verificar, um conjunto novo de conceitos e características, dentre as quais se destacam:

---

[6] http://www.orkut.com
[7] http://www.gmail.com
[8] http://www.blogger.com
[9] http://www.delicious.com

- Web, como plataforma para processar, produzir ou consumir informação;
- Canalização da inteligência coletiva e colaborativa, permitindo a qualquer usuário produzir e consumir informação de forma simples e direta;
- Modelos leves de programação, que podem ser facilmente manipulados e evitam contínuo ciclo de lançamento de software;
- Software independente do dispositivo.

Alguns itens, que serão apresentados a seguir, destacam-se como elementos que evidenciaram a consolidação da Web 2.0 como plataforma de interação.

3.1.1 Interfaces Ricas

Um dos grandes diferenciais da Web 2.0 é a maneira colaborativa em que o usuário se posiciona em relação à Internet, porém todo esse trabalho envolvendo a criação de uma nova cultura só foi possível por uma mudança de estrutura em relação ao desenvolvimento tecnológico por que passaram os sistemas disponíveis na Web.

A tarefa de desenvolvimento tecnológico da plataforma ficou a cargo de especialista em desenvolvimento de sistemas para Web, que iniciou um processo de atualização das aplicações para Web, de forma que tivessem aparência e funcionalidades muito parecidas com os sistemas denominados desktop. A principal modificação aparente para o usuário foi decretada a partir do momento em que as interfaces passaram a processar as informações solicitadas pelo usuário, sem a necessidade de atualização da página a qual o usuário estava conectado.

Essa novas interfaces, denominadas ricas, que passaram a ocupar grande parte dos sites, caracterizam-se pelo uso de um conjunto de tecnologias denominado Ajax, acrônimo de Asynchronous JavaScript and XML. Ajax não é uma nova tecnologia, mas sim uma técnica que reúne um conjunto de tecnologias, de forma que possa fornecer funcionalidades de desktop aos sistemas Web.

A tecnologia Ajax é, tecnicamente, a grande responsável pela forma dinâmica e rica que os aplicativos Web têm se apresentado e, portanto, tem sido fundamental para incentivar que, cada vez mais, novos usuários se aproximem da proposta da Web colaborativa.

Segundo Kalback (2007, p.345),

> Tecnicamente, uma aplicação web é um recurso em um site que realiza uma função. Uma pesquisa (busca) do site é uma aplicação web. Assim também, é um carrinho de compras ou o processo de pagamento em um site de comércio eletrônico. Mas estes são exemplos simples. Aplicações ricas são referidas como application rich (RIAs), no entanto, são uma classe de aplicações web mais sofisticadas que se comportam do mesmo modo que programas de software para desktop. Comparadas às páginas web tradicionais, elas são ricas em interação, ricas em conteúdo e ricas em funcionalidades. (tradução nossa)

Assim, evidencia-se que as interfaces ricas são a forma em que as informações são apresentadas tecnicamente, ou seja, o comportamento da interface em relação à interação do usuário com a mesma.

3.1.2 Inteligência Coletiva

O elemento mais característico da Web 2.0, e que pode ser aplicado em todos os outros conceitos, recursos e técnicas apresentadas por esse novo paradigma, é certamente a inteligência coletiva.

A inteligência coletiva abrange o conceito de comunidades, redes sociais, colaboração e discussão. A comunicação exercida pelas pessoas faz com que seja construída uma estrutura de aprendizado e de criatividade.

O termo inteligência coletiva põe diante de quem o observa a união de duas significativas palavras: inteligência e coletiva. A palavra inteligência, ao ser ouvida pode, sem muito esforço, levar o indivíduo a pensar a respeito de tudo o que se encontra armazenado em sua cabeça desde o dia de seu nascimento até àquele exato momento; conseqüentemente, também o leva a se questionar se é ou não inteligente (LEVY, 1999).

No pensamento de Pierre Lévy (1999, p. 28), se vê com clareza que inteligência coletiva "é uma inteligência distribuída por toda parte, incessantemente valorizada, coordenada em tempo real, que resulta uma mobilização efetiva das competências".

A construção de informação através das redes sociais, dos wikis e dos blogs estabelece uma rede participativa e interativa de comunicação, unindo, em algumas situações, pessoas com as mesmas características, dado o tipo de assunto abordado, e, em outras situações, pessoas de características totalmente diferentes.

O uso e a participação de pessoas com idades, sexo, formação e outras características diferentes consolidam e contribuem para o melhor desenvolvimento da plataforma em uso. Basta observar que alguns serviços, como Orkut, Flickr, Delicious, YouTube, entre outros, têm apresentado sensíveis mudanças no decorrer dos anos, fruto do processo colaborativo nas sugestões de desenvolvimento da plataforma.

Neste novo cenário, os usuários passam a ter participação ativa, porque produzem, criticam, alteram e sugerem novos conteúdos, deixando de ser simples telespectadores e passando a ser, além de consumidores, fornecedores de informação.

3.1.3 Wikis e Blogs

O termo Wiki, cunhado por Ward Cunninghan, autor do primeiro wiki, em 1995, foi inspirado na palavra wiki-wiki (super-rápido) da língua

havaiana. O objetivo inicial de Cunnighan era desenvolver um site que desse aos usuários cadastrados o acesso a conteúdos, permitindo alterar, gerenciar, criar novos conteúdos e disseminar as informações ali publicadas.

O modelo de Cunnighan tornou-se um padrão de desenvolvimento de conteúdo colaborativo, principalmente após o surgimento da Wikipédia, enciclopédia colaborativa multilíngüe (SCHONS, SILVA e MOLOSSI, 2007).

O que distingue o sistema Wiki de outras páginas da Internet é que o conteúdo pode ser editado e atualizado pelos usuários constantemente, sem necessidade de autorização do autor da versão anterior. Este sistema permite corrigir erros e inserir novas informações, ou seja, ninguém é autor proprietário de nenhum texto e o seu conteúdo é atualizado porque pode ser reformulado. Assim, wikis são sites que, além visualizados, pesquisados e terem conteúdos adicionados, podem ser editados diretamente por qualquer pessoa (RUPLEY, 2003).

Segundo a própria Wikipédia (2009),

> Wikipédia é uma enciclopédia multilíngue online livre colaborativa, ou seja, escrita internacionalmente por várias pessoas comuns de diversas regiões do mundo, todas elas voluntárias. Por ser livre, entende-se que qualquer artigo dessa obra pode ser transcrito, modificado e ampliado, desde que preservados os direitos de cópia e modificações, visto que o conteúdo da Wikipédia está sob a licença GNU/FDL (ou GFDL) e a Creative Commons Attribution-ShareAlike. Foi criada em 15 de Janeiro de 2001.

O rápido desenvolvimento e sucesso da Wikipédia[10] impulsionou o uso da ferramenta Wiki, de forma geral. Atualmente, é comum verificar que instituições de vários segmentos mantêm uma ferramenta Wiki internamente, para que seus funcionários e colaboradores possam construir conhecimento de forma coletiva.

Os blogs também se caracterizam como ambientes de sucesso, principalmente por passarem a oferecer um canal de comunicação direto

[10] http://pt.wikipedia.org

entre pessoas, sejam elas ligadas a empresas, governos, ou simplesmente poetas da informação pessoal ou cultural.

Os serviços de blog variam bastante, mas têm sempre a mesma característica, a de ser um ambiente aberto, que permite ao usuário postar informações sempre que desejar. Dependendo do contexto e da maneira que as informações são abordadas, estas ferramentas transformam pessoas comuns em celebridades.

É possível encontrar algumas outras variantes oriundas dos blogs, como os fotologs, que têm como característica principal a postagem de fotos, ou seja, o usuário deixa de oferecer seu álbum de fotografias para quem visita a sua casa e passa a disponibilizá-lo abertamente ao mundo.

Atualmente, uma das variações de serviço de blog que mais vêm despertando atenção dos internautas é o serviço de microblog Twitter, responsável por permitir pequenas postagens de no máximo 140 caracteres, onde os usuários podem "seguir" a postagens dos usuários que desejarem. Os 140 caracteres que delimitam as mensagens postadas no microblog Twitter foram definidos no tamanho da mensagem SMS de celulares.

Figura 5 – Twitter do Governador do Estado de São Paulo – José Serra
Fonte: http://twitter.com/joseserra_

Hoje, o Twitter é responsável por publicar informações mais rapidamente que outros meios de comunicação, como TV e mesmo os portais

de informações na Internet, e tem se destacado por apresentar os mais variados tipos de assunto e usuários. Exemplos como: a padaria do Supermercado Farinha Pura, do Rio de Janeiro, que avisa seus clientes através da mensagem “Saindo pãozinho agora”, e que virou rotina para os moradores da região; e de pessoas populares, como o governador José Serra, que mantêm contato com a comunidade, dando informações sobre medidas do governo e também sobre gostos e rotinas pessoais (figura 5), são apenas alguns exemplos de como a Web 2.0 tem passado a fazer parte da vida das pessoas, de uma forma geral.

3.1.4 Mashup

Utilizar o conceito de mashup em uma aplicação está diretamente relacionado a utilizar conteúdo de mais de uma fonte para criar novos serviços.

O mashup sistematiza uma interação de modo que as aplicações são quebradas em componentes de serviços, que, por sua vez, podem ser combinados e misturados com outros serviços, de acordo com as necessidades do negócio. Ambos permitem a reutilização de informações e de serviços já disponíveis para a criação de novas aplicações sob medida para o usuário. Este conceito envolve a disponibilização dos serviços através de APIs, pois elas fornecem acesso dinâmico a dados disponibilizados por vários sites ao mesmo tempo.

Os principais itens que caracterizam a utilização de mashups são:

- Uso de linguagem e plataformas padronizadas como HTML, XHTML e Javascript;
- Consumo de WebServices;
- Combinação de diferentes fontes, produzindo um conjunto de informações.

A sistematização do uso de mashups está diretamente ligada à possibilidade de agregar vários serviços em apenas um local na Web, como, por exemplo, utilizar dentro de um site a ferramenta Google Maps da Google, ou então o serviço de envio de mensagens das operadoras de telefonia móvel.

A utilização de mashup fortalece a agregação de valor a um ambiente informacional, de forma que facilita e contribui com o desenvolvimento de um ambiente com a utilização de ferramentas que estão disponíveis para uso aberto.

3.1.5 RSS (Really Simple Syndication)

O serviço de RSS, um dos principais serviços, entre os itens que compõem as funcionalidades da Web 2.0, é constituído por um conjunto de regras em XML, que permitem que os usuários publiquem informações ou as consumam diretamente de um site, sem precisar acessá-lo.

O formato RSS especifica o conteúdo XML de um noticiário. Alguns sites oferecem o serviço com o nome de “RSS Feed” ou ainda “Web Feed”.

A agregação de RSS funciona como um serviço de recebimento de mensagens através de um software, coletando apenas o cabeçalho das notícias e informações das mais variadas fontes. Atualmente, alguns sites oferecem agregadores de forma online, na própria Web.

Figura 6 - Canais RSS – Terra

Fonte: http://www.terra.com.br/rss/

Conforme se observa na figura 6, os principais portais de informações oferecem canais de RSS para que o usuário possa desfrutar do serviço de forma individualizada, ou seja, pode escolher o tipo de informação que deseja receber. Apesar de ter iniciado com os grandes portais de informações, hoje em dia, o serviço de RSS passou a ser amplamente utilizado e é possível receber informações das mais variadas fontes através dos “feeds”, além da forma mais tradicional que são as notícias, como programação de canais de TV, novas postagens em blogs e lançamento de novidades em sites de e-commerce.

Segundo Almeida (2007, p.2),

> [...] trata-se de uma tecnologia emergente, popularizada pelo conjunto de formatos padronizados, por meio do qual é possível oferecer aos usuários notificações automáticas sobre a atualização de conteúdos disponibilizados sob a plataforma Web.

Atualmente, a tecnologia é tão popular que alguns usuários consideram estranho acessar portais e sites que não disponibilizem o recurso.

3.1.6 Folksonomia

Folksonomia é a tradução do termo criado por Thomas Vander Wal, a partir da junção das palavras folk (povo) com taxonomy (Taxonomia). Wal (2006, p.1) define Folksonomia como "resultado de atribuição livre e pessoal de tags (etiquetas) a informações ou objetos (recursos na web), visando a sua recuperação".

Entre os recursos apresentados até então como funcionalidades da Web 2.0, a Folksonomia é um dos que mais caracterizam essa condição, de construção coletiva de inteligência informacional.

No capítulo 6 será abordado, de forma mais aplicada, o uso de Folksonomia, justamente porque é considerada elemento fundamental no desenvolvimento desta pesquisa, funcionando como recurso primordial na construção do modelo Representação Iterativa. Portanto, este tópico tem a característica apenas de definir o conceito e a aplicação da Folksonomia como funcionalidade da Web 2.0.

O conceito de Folksonomia também remete a estudos sobre taxonomia, e consequentemente a vocabulários controlados, que são instrumentos importantes na construção do conhecimento. Esses conceitos serão apresentados no próximo capítulo.

Golder e Huberman (2006, p.199) registram:

> A principal diferença técnica de uma folksonomia para uma taxonomia é que a primeira não estabelece uma relação hierárquica entre as classes (no caso, as tags), nem exige exclusividade entre as classes (um elemento pode pertencer a mais de uma classe).

O propósito principal da Folksonomia é permitir que usuários comuns criem labels/tags que possam descrever ou apontar para o conteúdo que estão inserindo na Internet, de modo que os recursos possam ser recuperados posteriormente pelo próprio usuário ou ainda por outros usuários que procurem informações no ambiente digital em que as

informações foram inseridas. Alguns serviços e sites, como YouTube, Delicious, Wordpress e Flickr, oferecem esse recurso.

Segundo Silva e Silva (2009, p. 202),

> O três pivôs da folksonomia são: o usuário (tagger), o objeto e a tag. Uma folksonomia tem seu alicerce centrado na tag, que é o elemento de classificação para o objeto, dessa forma, uma atenção especial deve ser direcionada ao uso de termos (tags) em uma categorização.

Várias são as definições apresentadas para descrever o conceito principal de Folksonomia. Entende-se que ela se caracteriza como uma forma de inserir e relacionar recursos através da descrição dos mesmos pelas palavras-chave, de forma aberta, que tem como principal objetivo facilitar o processo de gerenciamento e recuperação das informações em ambientes digitais.

Guy e Tonkin (2006, p.1) afirmam que,

> as etiquetas são apenas um tipo de metadados e não são um substituto para os sistemas de classificação formal como Dublin Core, MODS, etc... Ao contrário, elas são um meio suplementar para organizar as informações e ordenar os resultados de pesquisa (tradução nossa).

As tags podem ser definidas ainda como palavras-chave, categorias ou metadados, e podem podem ser classificados como qualquer palavra que define uma relação entre o recurso on-line e um conceito na mente do usuário (GUY e TONKIN, 2006).

Catarino (2009, p.46) define Folksonomia como

> [...] resultado da etiquetagem de recursos da Web num ambiente social (compartilhado e aberto) pelos próprios utilizadores da informação visando a sua recuperação. Destacam-se três fatores essenciais: 1) é resultado de uma indexação livre do próprio utilizador do recurso; 2) objetiva a recuperação a posteriori da informação; 3) É desenvolvida num ambiente aberto que possibilita o compartilhamento e até, em alguns casos, a sua construção conjunta.

Catarino tem uma visão mais social, do ponto de vista da descrição do recurso, porém também encontramos a relação de Folksonomia com categorização. De acordo com Marlow et. al (2006, p.1),

> [...] os sistemas que incorporam a folksonomia em seu funcionamento são chamados de Tagging Social. Para os autores, a prática de "etiquetar" um recurso é semelhante à categorização de bookmarks ("favoritos"). Não é à toa que se fala em Bookmarking Social, que são ferramentas que consistem no armazenamento de bookmarks em serviços online, os tagging systems.

O fato de a Folksonomia promover a participação do usuário de forma livre permite que a criação das tags receba o nome de vocabulário descontrolado, em uma alusão aos vocabulários controlados, que são um recurso disponível para alinhar indexação de informação dentro de um conjunto de palavras fixas que representam um determinado domínio de informação.

Aquino (2007, p.10) faz essa abordagem:

> Poderíamos dizer que a folksonomia é uma espécie de vocabulário descontrolado. Isso não quer dizer que o esquema seja uma desordem total [...]
>
> [...]Na verdade, trata-se de um mecanismo de representação, organização e recuperação de informações que não é feito por especialistas anônimos, o que muitas vezes pode limitar a busca por não trazer determinadas palavras-chave, mas sim um modo onde os próprios indivíduos que buscam informação na rede ficam livres para representá-la, organizá-la e recuperá-la, realizando estas ações com base no senso comum e tendo assim um novo leque de opções ao efetuar uma pesquisa para encontrar algum dado.

É possível verificar que a Folksonomia é um importante recurso em ambientes digitais de informação. E fica claro que são mais um recurso, e não, um recurso que venha substituir outros que já existem.

Guy e Tonkin (2006, p.1) afirmam que

> Concordamos com a premissa de que as tags não são substitutos para os sistemas formais, mas vemos isso como sendo a qualidade do núcleo que faz folksonomy tão útil.

É possível encontrar também quem dê o nome de Tag Clouds ao recurso de Folksonomia, porém percebe-se que este nome é mais utilizado quando há referência ao recurso técnico do uso de Folksonomia. O nome Tag Clouds é principalmente utilizado em ambientes que não têm a característica de fundamentar o uso de social tags, mas sim de apresentar o recurso ao usuário como mais um "recurso informático" de recuperação de informações.

Figura 7 - Tag Clouds
Fonte: http://www.geek.com.br/

O nome Tag Clouds foi dado porque, em grande parte dos ambientes que usam esse recurso, a lista de palavras mais utilizadas e mais citadas está espalhada em uma área da tela, como se fosse realmente uma nuvem de palavras (figura 7).

Há ainda alguns autores que entendem a Folksonomia como um recurso de classificação, caso, por exemplo, de Guy e Tonkin (2006, p.1), que a definem como "um tipo de sistema de classificação distribuída, criada por um grupo de indivíduos, tipicamente os utilizadores do recurso. Os usuários adicionam tags para itens como imagens, vídeos, marcadores e texto".

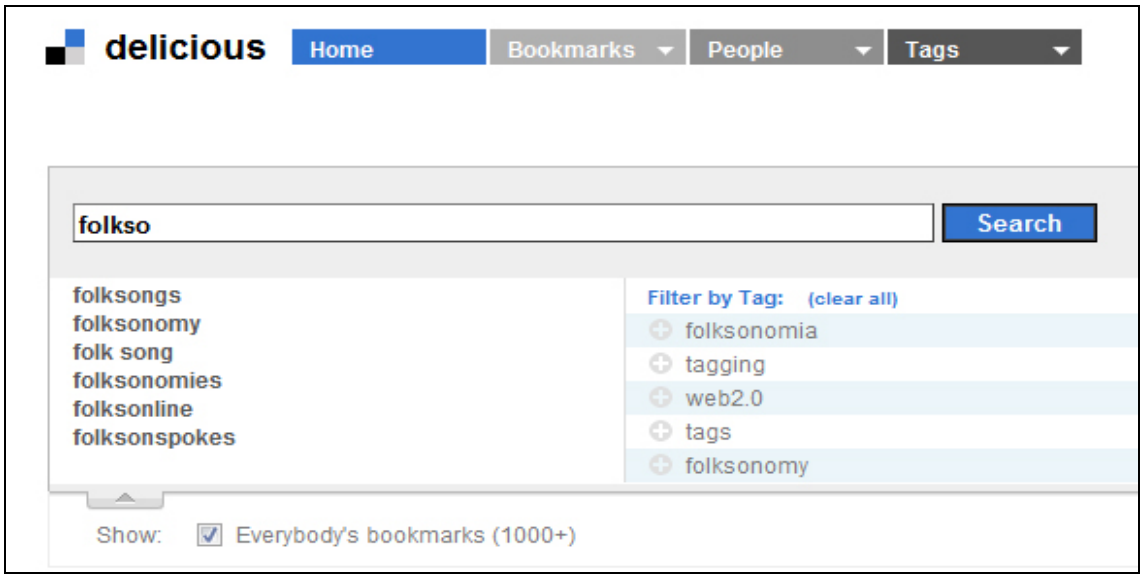


Figura 8 - Busca Del.icio.us
Fonte: http://delicious.com/search

A Folksonomia mudou o paradigma em relação à recuperação da informação em ambientes Web, tanto que é comum ver sites apresentando buscas baseadas em palavras-chave que foram inseridas pelo próprio usuário dentro do ambiente. Algumas ferramentas oferecem o serviço, mesmo sem creditar o conceito de inteligência coletiva neste contexto.

A funcionalidade tem recebido diversas adaptações, sendo que alguns sites fazem dessa característica seu principal ponto de apoio, como, por exemplo, do Del.icio.us (figura 8), que, conforme o usuário vai digitando a palavra a ser buscada, o próprio site vai sugerindo um conjunto de palavras, que têm a mesma grafia e que já foram amplamente utilizadas por outros usuários dentro do ambientes. Essa característica foi inicialmente apresentada através do Google Suggest e representa um facilitador ao usuário no momento da busca e descrição do recurso.

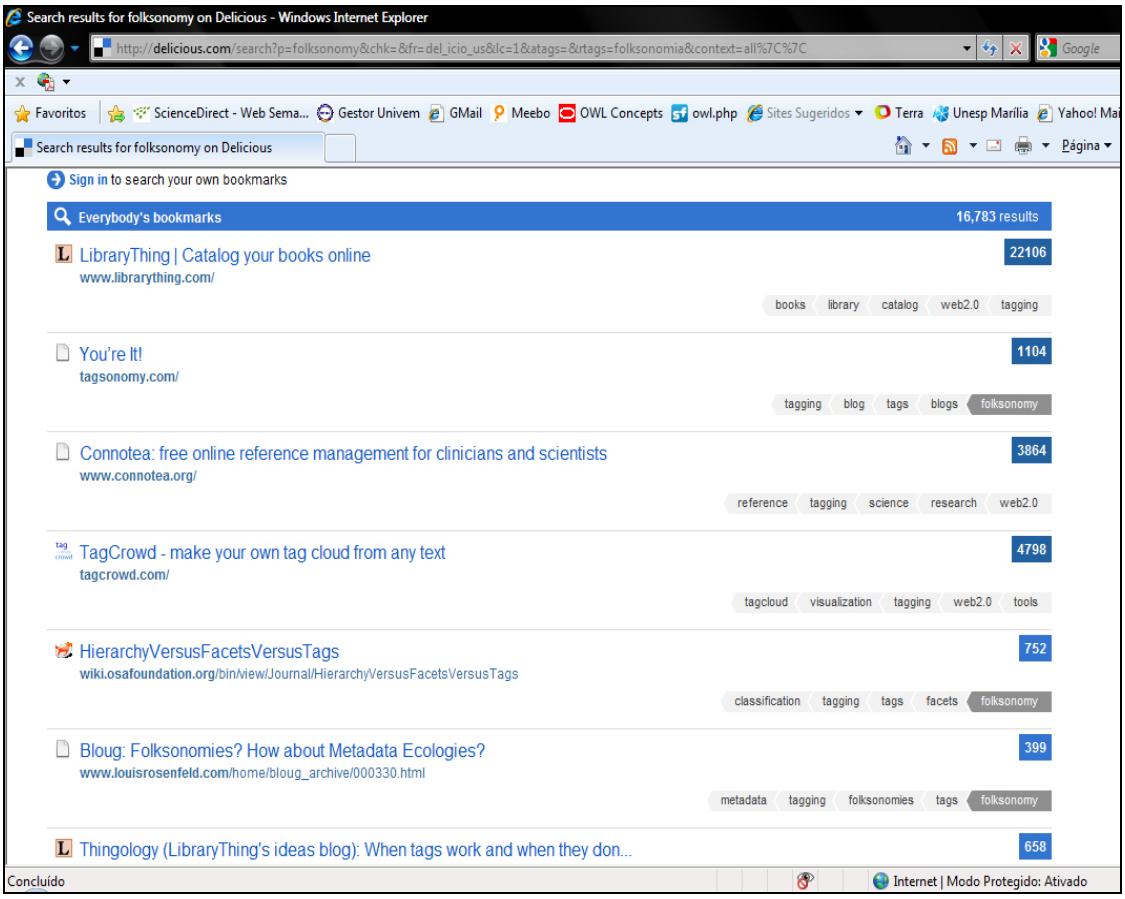


Figura 9 - Del.icio.us

Fonte: http://delicious.com/search?p=folksonomy

Outra adaptação, também operacionalizada no Del.icio.us (figura 9), é que a recuperação da informação é sempre baseada na palavra-chave utilizada pelo usuário, porém a ferramenta apresenta uma característica interessante de apresentar os resultados, que são os recursos cadastrados com a lista de palavras-chave, utilizadas no recurso no momento do cadastro, logo abaixo do link, facilitando o processo de busca por tags, com um simples clique em uma das palavras da lista, submetendo a nova recuperação de informação, baseada na palavra selecionada.

A Folksonomia é um recurso rico, que contribui de forma acentuada para o fortalecimento e solidificação da Internet como plataforma para construção de informação coletiva.

3.2 Web 3.0 – A Web Semântica

Web 3.0 é o termo que foi apresentado em 2006, pelo jornalista John Markoff, para se referir à terceira geração da Web. Os conceitos utilizados por John Markoff para cunhar o termo acabaram associando o nome a um termo já cunhado e utilizado anteriormente por Tim Berners-Lee, a Web Semântica, identificada como segunda geração da Web (PATRIOTA e PIMENTA, 2008).

Acredita-se que o termo Web 3.0 destaca algumas poucas novidades em relação à Web Semântica, porém acredita-se que, basicamente, os dois termos representam os mesmos princípios, que são de estruturar o conteúdo das informações a partir de conceitos semânticos, e é desta forma que também se entende nesta pesquisa.

A criação do projeto da Web Semântica, de Tim Berners-Lee, surgiu em face das dificuldades de localização, descrição e recuperação de informações em ambientes Web.

Um caminho para a solução da qualidade na recuperação dos dados que permita ao usuário resultados mais precisos parece ser a criação da Web Semântica, um projeto que visa dispor nos sites tanto informações descritivas e temáticas para os usuários, como informações que possam ser processadas e identificadas pelos computadores automaticamente. Assim, seria uma forma de disponibilizar informações para as máquinas/softwares juntamente com as informações para os usuários (BERNERS-LEE, LASSILA, HENDLER, 2001).

A Web Semântica trará uma estrutura ao significado da página Web, criando um ambiente propício para que os agentes de busca possam realizar tarefas sofisticadas e entregá-las ao usuário (BERNERS-LEE, LASSILA, HENDLER, 2001).

O desafio da Web Semântica vem sendo, a cada dia, prover uma linguagem capaz de expressar ao mesmo tempo dados e regras, de forma a

possibilitar a dedução de novos dados e regras a partir de qualquer sistema de representação de conhecimento a ser importado ou exportado na Web.

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

Essa separação permitirá uma recuperação da informação de várias maneiras, independente de como seja a busca, bastando que se conheça a estrutura dos dados. Este novo formato de recuperação de informação deverá facilitar a associação entre informações e ajudará a minimizar o problema da utilização de uma mesma informação em vários sistemas.

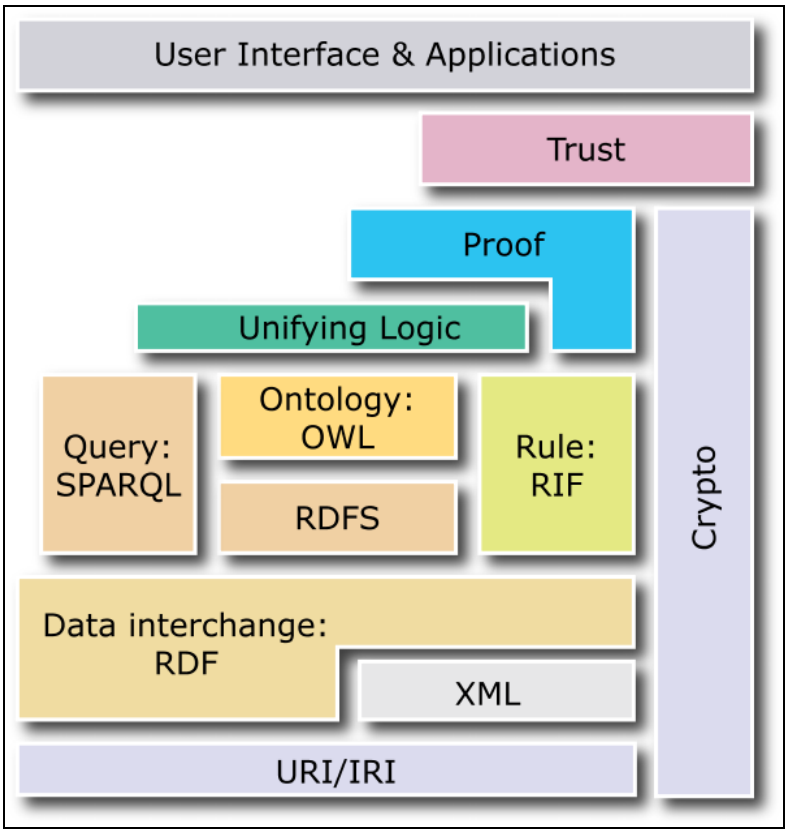


Figura 10 - Estrutura da Web Semântica (Layercake).
Fonte: http://www.w3.org/2007/03/layerCake.png

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

A proposta de Web Semântica delineada por Berners-Lee está representada na figura 10, onde são apresentadas as estruturas de camadas em que a Web Semântica está fundamentada.

Na camada base da figura 10 encontram-se URI (Uniform Resource Indentifiers) / IRI, que são os padrões para descrição de identificadores universais de recursos e códigos internacionais de dados. A camada denominada URI / IRI fornece a interoperabilidade em relação à codificação de caracteres e ao endereçamento e nomeação de recursos da Web Semântica.

O URI é um padrão para identificar um recurso físico ou abstrato de maneira única e global. Um identificador URL é um caso específico de URI, formado pela concatenação de sequências de caracteres para identificar o protocolo de acesso ao recurso, o endereço da máquina na qual o recurso pode ser encontrado e o próprio recurso em questão.

Para se entender melhor a parte da segunda camada nomeada XML, utiliza-se a seguinte citação de Greenberg (2003, p.6):

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

XML e XML Schema fornece a interoperabilidade em relação à sintaxe de descrição de recursos da Web Semântica. A Extensible Markup Language (XML) é uma linguagem para representação sintática de recursos de maneira independente de plataforma.

Os documentos que têm sua estrutura e seu conteúdo representados na linguagem XML são denominados de documentos XML. A XML Schema é uma linguagem de definição para descrever uma gramática (ou esquema) para uma classe de documentos XML. A linguagem XML Schema fornece elementos para descrever a estrutura e restringir o conteúdo de documentos XML. Os espaços de nomes (namespaces) fornecem um método para qualificar os nomes de elementos e atributos, utilizados nos documentos XML, através da associação destes nomes com os espaços de nomes identificados por referências de URI. Os espaços de nomes são úteis para distinguir entre dois elementos definidos com um mesmo nome, mas que pertencem a esquemas diferentes. Além disso, um documento pode associar elementos previamente definidos a sua estrutura, desde que utilize referências aos esquemas que definem esses elementos.

Segundo W3 Consortium (2009, p.2),

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

Dessa forma a linguagem XML se caracteriza como elemento facilitar no processo de processamento da informação.

Para Bax (2001, p.37),

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

A XML permite agregar semântica aos documentos, deixando por conta de cada aplicação a interpretação da marcação atribuída a este conteúdo. Esta abordagem amplia significativamente as possibilidades do uso das linguagens de marcação, entre elas a capacidade de definir

metadados – dados que descrevem dados. (CAMPOS; SANTACHE; TEIXEIRA, 1999)

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

A camada denominada RDF fornece um framework para representar informação (metadados) sobre recursos. As principais especificações do Resource Description Framework (RDF) abrangem um modelo de dados (para expressar declarações sobre os recursos), uma sintaxe baseada na Extensible Markup Language (XML) (para o intercâmbio das declarações) e uma linguagem de definição de esquemas para vocabulários.

A camada que apresenta ontologia com OWL, Rule: RIF, linguagem de consulta Sparql e RDFS fornece suporte para a evolução de vocabulários e para processar e integrar a informação existente, sem problemas de indefinição ou conflito de terminologia. A linguagem RDFSchema permite a construção de ontologias com expressividade e inferência limitadas, pois fornece um conjunto básico de elementos para a modelagem, e poucos desses elementos podem ser utilizados para inferência.

A Web Ontology Language (OWL) estende o vocabulário da RDF Schema para a inclusão de elementos com maior poder com relação à expressividade e inferência. Além disso, a linguagem OWL fornece três sub-linguagens para permitir o uso da linguagem por aplicações com diferentes requisitos de expressividade e inferência. O desenvolvedor pode escolher o módulo OWL adequado, de acordo com os requisitos da sua aplicação.

O principal conceito para associar informações é o uso de ontologias, pois através delas é possível representar ligações entre informações que sintaticamente não fazem nenhum sentido, porém

semanticamente têm conteúdos que estão direta ou indiretamente relacionados.

Por se considerar a abordagem sobre ontologias de fundamental importância, dedicar-se-á o próximo capitulo a este assunto, tratando dos principais quesitos para construção e manipulação de ontologias.

A linguagem RIF tem como objetivo principal fornecer suporte ao intercâmbio das diversas tecnologias baseadas em regras, para construção de ontologias.

Ainda completando a camada, temos a linguagem Sparql (Query Language for RDF), linguagem de consulta de informação que atua na recuperação de informação nos mais diversos tipos de estrutura de informação para Web Semântica, como RDF e OWL.

A camada denominada Lógica fornece suporte para a descrição de regras que expressem relações sobre os conceitos de uma ontologia, as quais não podem ser expressas com a linguagem de ontologia utilizada. As linguagens Rule Markup Language (RuleML) e Semantic Web Rule Language (SWRL) são exemplos de linguagens propostas para a descrição de regras para a Web Semântica.

As camadas denominadas Prova e Confiança fornecem o suporte para a execução das regras, alem de avaliar a correção e a confiabilidade dessa execução. Essas camadas estão em constante desenvolvimento e dependem muito da maturidade das camadas inferiores.

As iniciativas em torno da Web Semântica apontam para que o conteúdo disponível na Web seja codificado, de forma que seja possível o processamento automático pelos computadores. Desta forma, as pesquisas realizadas em mecanismos de busca, por mais complexas que sejam, retornariam apenas o resultado esperado, algo mais próximo, por exemplo, dos resultados apresentados por sistemas que têm informações armazenadas de forma estruturada. Para isso, é necessário padronizar um mecanismo consistente de metadados.

3.2.1 Metadados

Os documentos são mais fáceis de localizar e gerir se se conhecer algo sobre eles, como o nome do autor, data de publicação, assunto, etc. Esse tipo de informação, que define "dados sobre dados", é o conceito básico atribuído ao termo metadados. Ao disponibilizar um arquivo para download, um exemplo de metadados para este arquivo seria: nome do programa, versão, tamanho do arquivo, informações sobre a licença de uso, plataforma, etc.

Para que os recursos informacionais sejam recuperados em um sistema de informação (seja ele digital ou não), é preciso utilizar métodos de representação da informação para que ocorra a mediação entre a forma registrada (documento) e o usuário (PEREIRA e SANTOS, 1998).

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

Takahashi (2000, p.172) define metadados como

> Dados a respeito de outros dados, ou seja, qualquer dado usado para auxiliar na identificação, descrição e localização de informações. Trata-se em outras palavras, de dados estruturados que descrevem as características de um recurso de informação.

Fica evidente assim que metadados são informações a respeito da estrutura de outros dados.

Alves (2005, p. 115) apresenta uma definição mais completa sobre metadados.

> Metadados são conjuntos de atributos, mais especificamente dados referenciais, que representam o conteúdo informacional de um recurso que pode estar em meio eletrônico ou não. Já os formatos de metadados, também chamados de padrões de metadados, são estruturas padronizadas para a representação do conteúdo informacional que será representado pelo conjunto de dados-atributos (metadados). Em outras palavras, os formatos ou padrões de metadados podem ser considerados como formas de representação[6] de um item documentário.

Metadados são utilizados para descrever as características de recursos e seus relacionamentos. Tradicionalmente, o uso de metadados é associado a sistemas gerenciadores de banco de dados. Na última década, os metadados ganharam uma nova dimensão e adquiriram grande importância no gerenciamento e manutenção de data warehouses, mecanismos de busca, ferramentas de software, etc.

De forma geral, os metadados são um conjunto de informações que têm como característica principal reunir informações sobre a descrição de informações, ou seja, os metadados têm a função de armazenar um cabeçalho de informações que apresente os dados que estão sendo armazenados.

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

- Descrição de recursos: informação expressa através de metadados, determinado pelo objetivo e tipo do recurso.
- Produção de metadados: sumário da descrição dos dados, que pode se tornar um processo caro quando realizado manualmente. A tendência é realizar automaticamente esse processo, incentivado pelo uso das tecnologias XML e RDF.
- Uso de metadados: envolve o uso e acesso de metadados, é especialmente relevante para a localização de recursos na web. Neste contexto, os metadados devem incluir informações sobre os recursos, tais como a identificação, descrição, estrutura.

Portanto, na Web, o imenso conteúdo disponível e a heterogeneidade dos recursos evidenciam cada vez mais a necessidade de adoção de padrões para metadados, a fim de aprimorar e facilitar a recuperação da informação.

A criação de um único padrão de metadados que aborde todas as áreas do conhecimento humano seria o ideal, porém construir um padrão

que consiga abarcar toda a estrutura de informações e domínios de conhecimentos é uma tarefa de extrema complexidade.

Segundo Souza et al. (1997, p. 99), “os padrões de metadados têm como função fornecer as definições e formar uma rede para automatizar registros de propriedades e dados cadastrais de forma padronizada e consistente.”

Souza e Alvarenga (2004, p.5) afirmam:

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

Os padrões de metadados foram sendo desenvolvidos para diferentes finalidades: GILS (Government Information Locator Service), usado para descrever informações governamentais; FGDC (Federal Data Geographic Committee), usado na descrição de dados geoespaciais; MARC (Machine Readable Cataloging), usado para a catalogação bibliográfica; CIMI (Consortium for the Interchange of Museum Information), que descreve informações sobre museus.

Para localização de recursos na web, o padrão de metadados mais utilizado e difundido é o Dublin Core (DC), que apresenta uma estrutura a partir de um conjunto de descritores simples e genéricos que objetiva a descoberta e o gerenciamento de recursos na web. O Dublin Core não requer conhecimentos extremos de especialistas no momento de descrever os recursos, devido à simplicidade de utilização, podendo ser usado por qualquer tipo de usuário, característica evidenciada pelo W3C para recomendar seu uso como padrão de metadados para descrever recursos na Web.

3.2.2 Dublin Core

O padrão Dublin Core é uma iniciativa para criação estruturas de informação, para uso na Web, baseado no pressuposto de que a busca por recursos de informação deve ser independente do meio em que estão armazenados.

O padrão é atualmente mantido pela Dublin Core Metadata Initiative (DCMI), que teve início em 1995, ganhando o nome da localidade onde se deu o encontro inicial, Dublin, Ohio, USA.

Segundo Lagoze (1996, p.1), "o Dublin Core pretende ser simples e para facilitar o uso pelos criadores e mantenedores de documentos web, descritivo o suficiente para auxiliar na recuperação de recursos na Internet." (tradução nossa)

O DC foi inicialmente sugerido com 15 elementos, constituído de pares (nome atributo / valor atributo) que formam o núcleo principal do padrão, e é nomeado como forma simples, porém, devido a grande diversidade de utilização do padrão, constantemente o DCMI tem ampliado o padrão, ampliando as possibilidades de uso dos elementos através da adição de qualificadores.

Os 15 elementos iniciais do padrão são apresentados no quadro 1:

Elemento	Descrição	Comentário
contributor	Uma entidade responsável por fazer contribuições para o recurso	Exemplos de um contribuinte incluem uma pessoa, uma organização ou um serviço.
coverage	Indica onde o recurso está fisicamente localizado.	
creator	Pessoa ou organização responsável pelo conteúdo	Exemplos de um Criador incluem uma pessoa, uma organização ou um serviço. Normalmente, o nome de um Criador deve ser utilizado para indicar a entidade.
date	Data em que o recurso se tornou disponível.	

Elemento	Descrição	Comentário
description	Descrição do conteúdo	Descrição pode incluir, mas não está limitado a: um resumo, uma tabela de conteúdo, uma representação gráfica, ou um texto livre sobre do recurso.
format	O formato no qual o recurso se apresenta. Suporte físico ou dimensões do recurso.	Exemplos incluem tamanho e duração. Uma prática recomendada é utilizar a lista de Tipos de Mídia Internet [MIME]. http://www.iana.org/assignments/media-types/
identifier	Uma referência inequívoca para o recurso dentro de um determinado contexto, tal como uma URL.	
language	O idioma em que está escrito o recurso.	Melhor prática recomendada é a utilização de um vocabulário controlado.
publisher	Uma entidade responsável por tornar o recurso disponível	
relation	Como o conteúdo se relaciona com outros recursos, como, por exemplo, se é um capítulo em um livro	
rights	Um ponteiro ou link para uma nota de copyright	
source	Fonte de onde foi originado o conteúdo.	
subject	O assunto ou tópico coberto pelo documento	
title	Nome dado ao recurso ou título.	
type	Uma categoria preestabelecida para o conteúdo	

Quadro 1: Elementos básicos do DC

Os elementos extras que complementam os 15 elementos definidos pelo DC são denominados de qualifiers. Esses qualifiers são avaliados pelo

DCMI (Dublin Core Metadata Initiative) para fazerem parte do conjunto de descritores às aplicações.

Os qualifiers têm como objetivo principal estender e qualificar os descritores básicos.

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

Para a representação destes qualifiers, é dada preferência aos vocabulários, anotações formais e termos mantidos e estabelecidos pelas agências já conhecidas dos usuários. Os implementadores desenvolvem qualifiers adicionais para uso dentro de aplicações e domínios específicos. Tais qualifiers podem ser reusados por outras comunidades dentro do contexto mais amplo (DCMI, 2008).

Segundo o DCMI (2008), o Dublin Core Qualifiers possui duas classes:

- Refinamento do Elemento: um elemento refinado compartilha o significado do elemento de uma maneira mais específica e restrita. Se não compreender o refinamento do elemento, o usuário deve ignorar o qualifier e retornar ao elemento geral.
- Esquema de Codificação: identificam esquemas que auxiliam na interpretação de um elemento. Esses esquemas incluem vocabulários controlados e anotações formais ou regras para a representação do mesmo.

Através do quadro 2 é possível verificar os elementos com seus respectivos qualifiers.

Elemento	**Qualifier**	**Comentário**
Audience	Mediator	Uma entidade que intermedia o acesso aos recursos e para quem o recurso se

Elemento	Qualifier	Comentário
		destina ou é útil.
	EducationLevel	Nível de escolaridade ou formação para o qual o recurso foi escrito (destinado).
Title	Alternative	Uma alternativa para o nome do recurso.
Description	TableOfContents	A lista das subunidades do recurso
	Abstract	Resumo sobre o Recurso.
Date	Created	Data da criação do recurso.
	Valid	Data (muitas vezes um intervalo) de validade de um recurso.
	Avaliable	Data (muitas vezes um intervalo) de que o recurso está ou estará disponível.
	Issued	Data de emissão formal (por exemplo, a publicação) do recurso.
	Modified	Data que o recurso foi alterado.
	DateAccepted	Data de aceite do recurso.
	DateCopyrighted	Data de direitos autorais.
	DateSubmitted	Data de submissão do recurso.
Format	Extent	O tamanho ou a duração do recurso.
	Medium	O material ou estrutura física portadora do recurso
Relation	isVersionOf	Um dos recursos relacionados com o qual o recurso descrito é uma versão, edição ou adaptação.
	hasVersion	Um recurso que está relacionado com uma versão, edição ou adaptação dos recursos descritos.

Elemento	Qualifier	Comentário
	isReplacedBy	Recurso que pode substituir ou suplantar o recurso descrito.
	replaces	Um recurso que está relacionado suplantado, deslocado, ou substituído pelo recurso descrito.
	isRequiredBy	Um recurso que apóia a funcionalidade ou coerência do recurso descrito.
	requires	O recurso requer outro recurso para apoiar sua funcionalidade ou coerência.
	isPartOf	Recurso relacionado ao principal, ao qual está física ou logicamente incluído.
	hasPart	Um recurso que está incluído ou relacionado fisicamente ou logicamente no recurso principal descrito.
	isReferencedBy	Um recurso que as referências relacionadas, cita, ou aponta para o recurso descrito
	references	Um recurso que é referenciado ou então aponta para o recurso descrito.
	isFormatOf	Um recurso relacionado ao principal recurso descrito, mas em outro formato.
	hasFormat	Outra maneira de descrever o formato do recurso.
	conformsTo	O padrão estabelecido para o qual o recurso foi descrito.
Coverage	Spatial	Características espaciais do recurso

Elemento	Qualifier	Comentário
	Temporal	Características temporais do recurso
Rights	AccessRights	Informações sobre quem pode acessar o recurso ou uma indicação de seu status de segurança
Identifier	BibliographicCitation	Uma referência bibliográfica para o recurso.

Quadro 2: Qualificadores de elementos do padrão Dublin Core

Descrever o conteúdo e não apenas exibi-lo é o primeiro passo para a criação da Web Semântica. A utilização das técnicas e tecnologias apresentadas segundo a clássica figura da Web Semântica (figura 10) é de fundamental importância para a constituição de um ambiente baseado em recuperação de conteúdo.

3.2.3 Web Standards

Apesar dos principais browsers, entre outras tecnologias de acesso a Web, estarem diretamente envolvidos na criação dos padrões Web desde a formação do W3C, a utilização dos padrões na construção dos browser não tem sido efetiva. Ao lançar browsers que não suportam os padrões, os fabricantes fragmentam desnecessariamente a Web, prejudicando de igual forma designers, programadores, utilizadores e empresas.

A falta de suporte uniforme para os padrões do W3C acaba deixando usuários e programadores frustrados, porque não conseguem ter o mesmo resultado no acesso aos dados com qualquer browser que escolham.

Em resposta a estes problemas, o Web Standards Project (WaSP) foi formado em 1998 com o objetivo de promover os padrões Web e encorajar os fabricantes de browsers a fazer o mesmo, assegurando desse modo um

acesso simples e com menos custos para todos. (THE WEB STANDARDS PROJECT, 2009).

O desenvolvimento e, consequentemente, o uso destes padrões tendem a facilitar o trabalho de interoperabilidade entre as informações atualmente depositadas na Web, e podem ser verificadas no site do Projeto Web Standards[11].

Entre as propostas para se encaminhar a Web para um ambiente que possa interagir com os usuários, de modo a agilizar a coleta de informações, está a criação dos Web Standards, que são um conjunto de padrões produzidos pelo W3C e destinados a orientar fabricantes, desenvolvedores e projetistas para o uso de práticas que possibilitem a criação de uma Web acessível a todos, independentemente dos dispositivos usados ou de suas necessidades especiais.

As possibilidades criadas com o uso dos padrões têm a intenção principal de permitir que os sites desenvolvidos através do uso destes padrões possam ser interpretados em qualquer tipo de ambiente que tenha acesso a Web, como os próprios browsers nos mais variados formatos, versões e sistemas operacionais, assim como nos mais variados tipos de dispositivos móveis ou ainda em TVs digitais com tecnologia de acesso à Internet.

Os Web Standards podem ser divididos, de certa forma, em três principais partes, visto que elas representam uma sugestão de divisão real do conteúdo formal das páginas web.

A primeira parte cuida especificamente da parte estrutural do desenvolvimento web, onde estão relacionadas as informações sobre as principais partes de um documento web, além dos cuidados com a semântica e também com a composição das tags que formarão o documento web. Atualmente, os dois principais padrões de estrutura para desenvolvimento web são: HTML 4.01[12], que foi recomendado pelo W3C, a

[11] http://www.webstandards.org
[12] http://www.w3.org/TR/html4/

partir de 1999; e o XHTML 1.0[13], recomendado pelo W3C[14], a partir de 2000, com revisão em 2002. O padrão HTML 5.0 está atualmente em fase experimental.

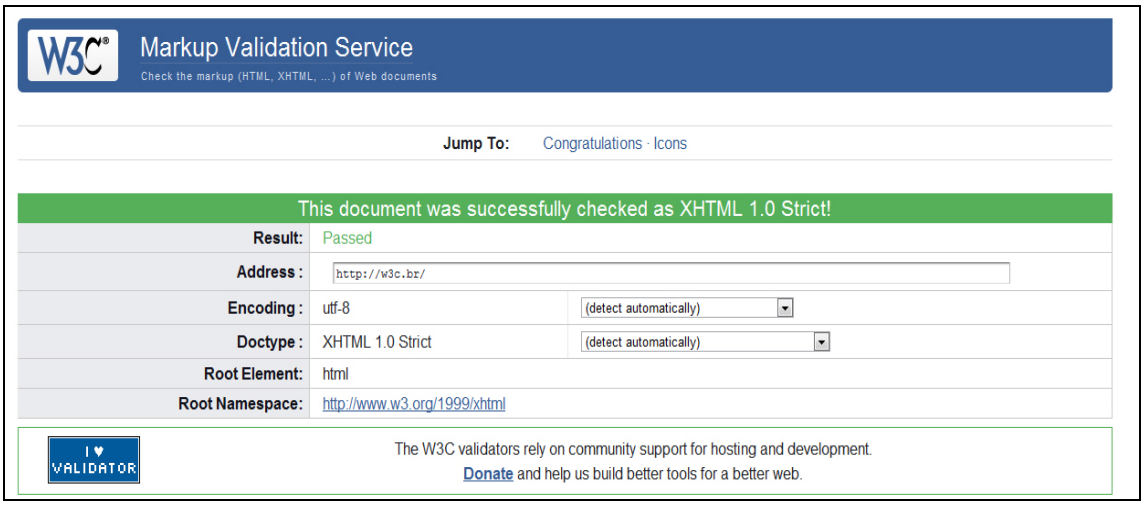


Figura 11 - Validação Web Standard do site da W3C Brasil
Fonte: http://validator.w3.org/

Apesar de os padrões terem sido recomendados há quase 10 anos, raramente se encontram sites que estão adequados ao padrão e que seguem rigorosamente as normas estabelecidas.

É possível validar se um determinado site ou página web está adequado ao padrão através do validador disponível no site do W3C.

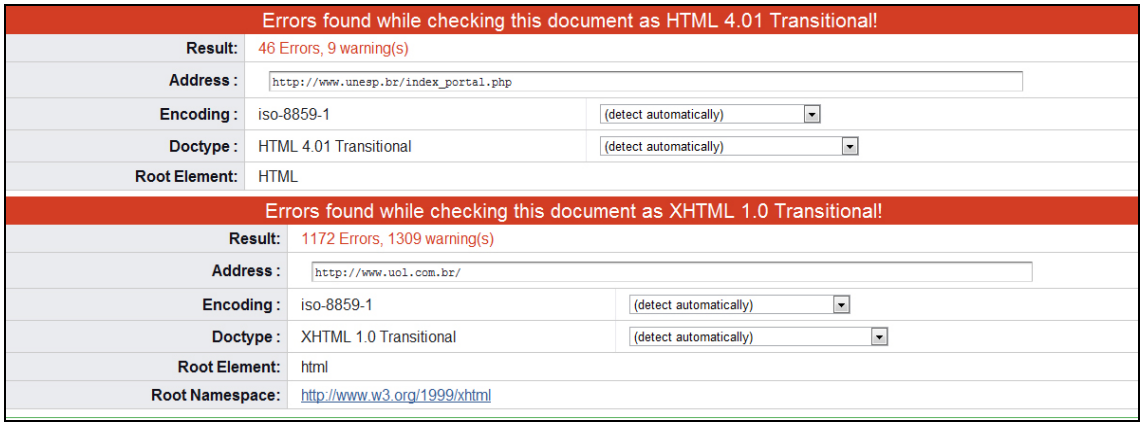


Figura 12 - Validação Web Standard dos portais Uol e Unesp
Fonte: http://validator.w3.org/

[13] http://www.w3.org/TR/xhtml1/
[14] http://www.w3.org/

Através das figuras 11 e 12, é possível verificar o teste realizado em três portais para validação dos padrões de desenvolvimento. Na figura 11, observa-se o teste do portal do W3C no Brasil, que passou com sucesso, e por isso é possível verificar no próprio site (figura 13) um selo de que o site está validado nos padrões do formato XHTML 1.0. Na figura 12, é possível ver o teste realizado no portal Unesp e no portal UOL, que apresenta uma gama de erros e problemas que o incompatibilizam com o padrão sugerido.

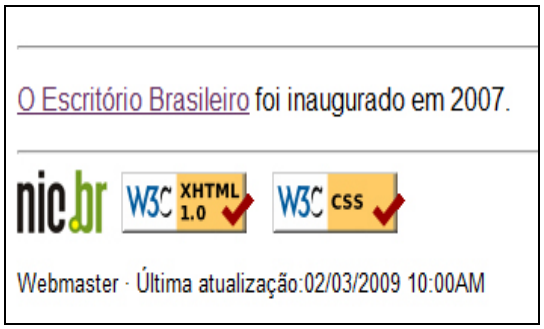


Figura 13 - Selo de validação Web Standard - padrão XHTML 1.0, no site do W3C Brasil

Fonte: http://w3c.br/

A segunda parte dos padrões Web Standards trata da questão da apresentação das informações, que compreende principalmente os aspetos visuais que não podem ser considerados informações textuais.

O padrão adotado para apresentação visual e recomendado pelo W3C é o CSS[15] – Cascading Style Sheets, que atualmente se encontra na versão CSS 2.1, recomendada pelo W3C, a partir de Abril de 2009.

A tecnologia CSS permite com que haja uma divisão clara entre estrutura e forma na composição de um site.

O W3C também apresenta um validador para verificar se a utilização do padrão CSS está correta, que pode ser encontrado no endereço: http://jigsaw.w3.org/css-validator/.

[15] http://www.w3.org/Style/CSS/

A terceira parte dos padrões Web Standards recai sobre a criação de efeitos e comportamentos que o site possa ter. Esses efeitos podem ser implementados para serem executados tanto do lado do cliente/usuário como do lado do servidor, e são implementados através de scripts de programação e recursos como utilização de Javascript e Ajax.

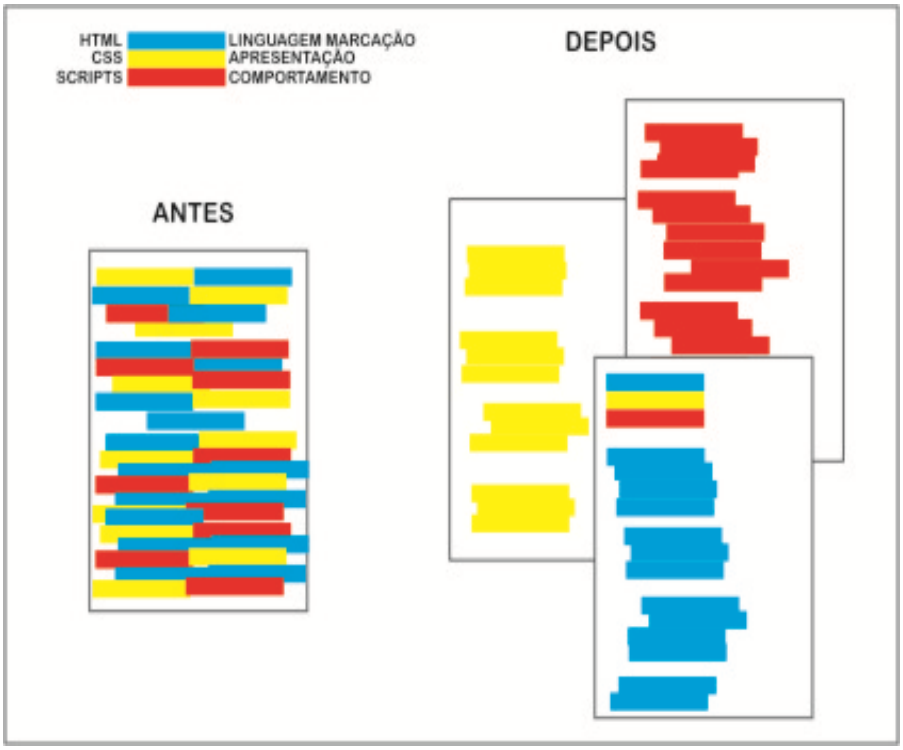


Figura 14 - Aplicação de Web Standards em um documento Web.
Fonte: Próprio autor

A aplicação dos padrões possibilita separar um site ou um documento web em três camadas distintas: estrutura, apresentação e comportamento (figura 14). De forma que fique muito mais fácil a manutenção do conjunto de informações, visto que nesse formato o portal ou site deixa de ter apenas um código unindo todas as informações, de forma misturada, para passar a ter códigos independentes para estrutura, apresentação e comportamento, de forma que essa estrutura fique transparente para o usuário, porém fique viável do ponto de vista de manutenção e apresentação, para o programador.

3.2.4 Microformatos

Com o surgimento e efetivação da Web 3.0 como um caminho a ser seguido no desenvolvimento de conteúdo para a Web, iniciaram-se as aplicações que, baseadas em alguns conceitos da Web 3.0, têm contribuído para que se possa separar estrutura de conteúdo e iniciar o processo de unir de forma semântica as informações.

Entre os tipos de aplicações desenvolvidas, destacam-se o uso de Microformatos, que são uma série de especificações, cujo foco principal é apresentar metainformações aos humanos e, posteriormente, às máquinas.

É uma nova maneira de se pensar sobre dados. Essa série de especificações constitui um “dicionário” de conteúdo semântico para (X)HTML, que tem como base os Web Standards e são escritas para descrever a informação da forma mais simples possível.

A principal função destas especificações é enriquecer a informação inserida em páginas web com metainformação, e isso é feito codificando os metadados no corpo do documento. O nome Microformatos está relacionado com a informação de "Pequenos formatos" (micro + format) de dados (informação) válidos no código do seu conteúdo XHTML.

Os Microformatos podem ocupar o lugar que antes era ocupado pela tags META do HTML. As tags META tinham o objetivo de apresentar metainformações sobre o conteúdo da página, sendo inclusive utilizados pelas ferramentas de busca para compor o banco de dados e, consequentemente, servindo de base para constituição do valor a ser recuperado. As tags META, devido ao abuso e mau uso (como forma de spam), passaram a ser desconsideradas pelas ferramentas de busca e caíram em desuso.

A diferença entre as tags META e os Microformatos é que, neste segundo, as informações são cadastradas no corpo do documento, diferente das tags META, que inseriam as metainformações no cabeçalho do documento.

Segundo Mendez, Bravo e Lópes (2007, p.109),

> [...] os Microformatos são apenas um conjunto de valores "especiais" ou finitos, utilizado para um fim específico. A particularidade destes valores é que normalmente faz parte de um determinado conjunto elementos que, por vezes, está associada a um padrão ou esquema (schema), amplamente adotada como hCard e hCalendar por exemplo", portanto o uso de Microformatos deve estar associado a um formato já descrito.

No site oficial dos Microformatos[16], é possível verificar as especificações/esquemas já estabelecidas como padrão para o uso de Microformatos, além de especificações que estão em processo de draft, e que devem vir a se tornar especificações recomendadas brevemente.

Entre os principais esquemas/especificações para uso de microformatos, destacam-se:

- Para pessoas e organizações: hCard e XFN (XHTML Friends Network);
- Calendários e eventos: hCalendar;
- Avaliação, classificação e opinião: hReview;
- Licenças: rel-license;
- Tags, palavras-chave e categorias: rel-tag;
- Listas e projetos: XOXO (Extensible Open XHTML Outlines)
- Entre as especificações em processo de recomendação, estão:
- adr – especificação que usa apenas o campo adr do hCard, tornando uma maneira simples de publicar a estrutura de um endereço Web.
- geo – especificação para marcação de coordenadas geográficas.
- hAtom - especificação para os conteúdos, em formato de feeds, que podem ser distribuídos, principalmente, mas não exclusivamente em weblog.

[16] http://microformats.org/

- hAudio – especificação para incorporação de informações sobre gravações em áudio.
- hMedia – especificação para informações sobre Imagens, Vídeos e Audios.
- hProduct – especificação para produtos e serviços na Web. Pode ser utilizado por serviços como de e-commerce, entre outros.
- hRecipe – especificação para receitas culinárias.
- hResume – especificação para resumos e currículos.
- hReview – especificação para opiniões sobre produtos, serviços, negócios, eventos, entre outros.
- rel-directory – especificação para indicar um diretório dentro de um hyperlink.
- rel-enclosure – especificação para indicar que um link representa um download de um arquivo
- rel-home – especificação para indicar um link para uma homepage
- rel-payment – especificação para indicar mecanismos de pagamento.
- robots exclusion – especificação para orientar os robôs (crawlers) quanto ao conteúdo que deve ou não ser indexado.
- xFolk – especificação para publicação de palavras-chave definidas pelos usuários. Baseado no conceito de Web 2.0, visto anteriormente.
- XFN - é uma maneira simples de representar as relações humanas usando hiperlinks. Pode ser utilizado em blogs para demonstrar relações.
- XOXO - é um formato de esboço simples, aberto escrito em padrão XHTML e adequado para ser embutido em (X)HTML,

> Atom, RSS, e XML arbitrário. XOXO é um dos muitos microformatos de padrões aberto.

Atualmente, os padrões de Microformatos mais difundidos são hCard, hCalendar e rel-tag. Esses esquemas tiveram aceitação rápida, porque foram os primeiros a ser desenvolvidos e permitiram que o usuário tivesse acesso a resultados através do seu uso. Os padrões hCard e hCalendar foram criados a partir dos já existentes padrões de Web, vCard e iCalendar, respectivamente.

Segundo Mendez, Bravo e Lópes (2007, p. 109),

> vCard e iCalendar são dois padrões de descrição e intercâmbio de informações, usadas em vários aplicativos e dispositivos como telefones celulares, PDA ou aplicações de PC (microcomputador). Usado para descrever cartões de visita com informações da instituição, endereço, telefone, e-mail, web site, etc. e para a descrição iCalendar eventos no tempo (compromissos, reuniões, conferências) com áreas específicas, tais como localização, data de início e no final, e assim por diante. Os padrões hCard e hCalendar deram mobilidade e versatilidade ao vCard e iCalendar no ambiente web, tornando-os microformatos adequados para inclusão em XHTML, Atom, RSS e mesmo xml.

O hCard é um formato baseado no formato vCard, para troca de informações de Address Book. Um vCard funciona como um cartão de visitas anexado as suas mensagens. Ele contém informações como o seu nome, email, endereço, telefone e site. Quando alguém recebe uma mensagem com o seu vCard, pode adicionar você ao catálogo de endereços, aproveitando todos os dados do vCard.

O hCard tem uma estrutura de informação muito parecida com a do vCard, e possibilita enviar informações pessoais ou de uma instituição através de uma página Web. Entre as informações que compõem a estrutura do hCard estão: endereço completo (com vários segmentos), email, coordenadas geográficas, apelido, foto, anotações, logotipo, entre outros.

```
<div id="hcard-José-Eduardo-Santarem-Segundo" class="vcard">
  <a class="url fn n" href="http://santaremsegundo.blogspot.com">
    <span class="given-name">José Eduardo</span>
```

```
        <span class="additional-name">Santarem</span>
        <span class="family-name">Segundo</span> </a>
        <div class="org">UNESP</div>
       <a class="email" href=mailto:santarem@marilia.unesp.br > santarem@marilia.unesp.br</a>
    <div class="adr">
        <div class="street-address">Rua Hygino Muzzi Filho, 737</div>
          <span class="locality">Marília</span> ,
          <span class="region">SP</span> ,
          <span class="postal-code">17515-420</span>
          <span class="country-name">Brasil</span>
        </div>
    </div>
```

EXEMPLO 3 - MICROFORMATO HCARD

O código apresentado no exemplo 3 é a estrutura de informação baseada no microformato hCard para definição dos dados de uma pessoa.

O site oficial dos Microformatos apresenta também uma ferramenta interativa para que o usuário possa criar um hCard sem a necessidade de conhecer o código de programação. Essa ferramenta, apresentada na figura 15, denominada hCard Creator (http://microformats.org/code/hcard/creator), cria automaticamente o hCard, baseado nos dados que o usuário cadastrar nos campos.

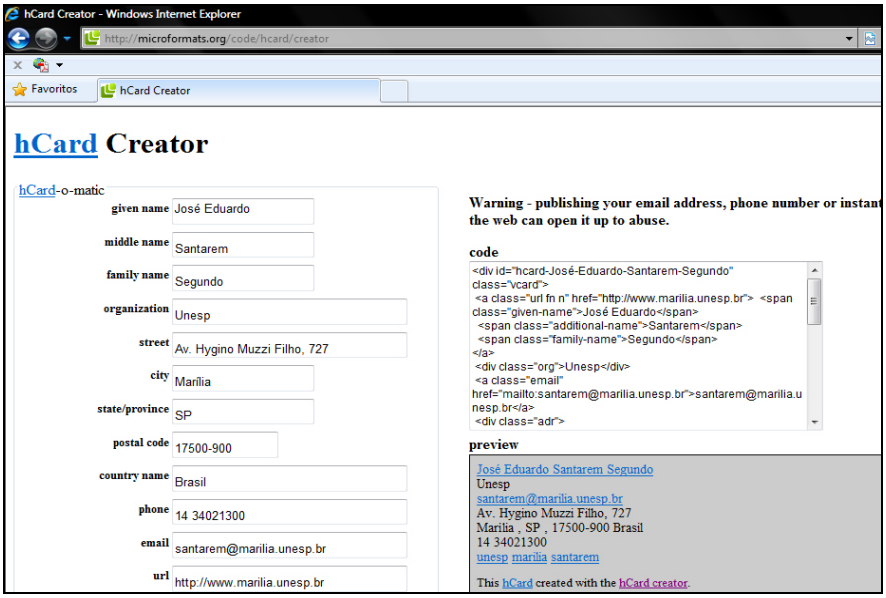


Figura 15 - hCreator
Fonte: http://microformats.org/code/hcard/creator

O uso de Microformatos embutidos na sua página Web insere em seu documento a estrutura da informação, de forma que até então não era possível fazer.

Outro Microformato consolidado é o hCalendar, baseado no padrão interoperável iCalendar, que integra informações sobre informações em determinada data.

O hCalendar armazena informações, como: data, resumo, local, duração, url, categoria, coordenadas geográficas, entre outros. Através deste Microformato, as páginas podem trocar informações diretamente sobre eventos, e permitir também que agentes recuperem essa informação de forma mais clara e ágil do que quando esse tipo de informação fica disponível apenas em texto puro, em formato HTML.

Para o formato hCalendar também há um hCalendar-Creator, disponível no site do projeto Microformatos[17], onde é possível criar a estrutura de informação de um evento em Microformato, para embutir no seu documento XHTML.

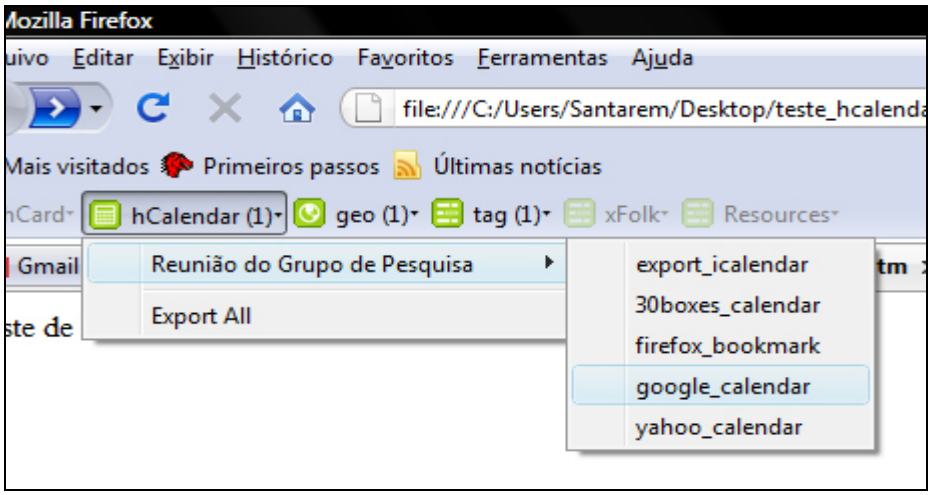


Figura 16 - Add-on Operator do Firefox identificando e disponibilizando informações sobre Microformato hCalendar

Fonte: Próprio autor

[17] http://microformats.org/code/hcalendar/creator

Apesar de os Microformatos já estarem disponíveis há algum tempo, ainda são poucas as ferramentas e agentes que exploram as páginas que contêm essas informações embutidas. Alguns browsers, como o Mozilla Firefox e o Opera, através de add-ons e extensions, deixam disponível ao usuário, se este desejar, a inclusão de agentes que detectam e permitem interatividade através dos Microformatos.

```
<div class="vevent" id="hcalendar-Reunião-do-Grupo-de-Pesquisa">
<a class="url" href="http://www.marilia.unesp.br">
  <abbr class="dtstart" title="2009-09-04T14:00-03:0000">September 4, 2009 2</abbr> –
  <abbr class="dtend" title="2009-09-04T17:00-03:00">5pm</abbr> :
    <span class="summary">Reunião do Grupo de Pesquisa</span>
    <span class="location">Unesp Marília</span></a>
    <div class="description">Reunião Introdutória do Grupo de Pesquisa de Novas Tecnologias da Informação</div>
    <div class="tags">Tags: <a rel="tag" href="http://eventful.com/events/tags/gpnti;unesp">gpnti;unesp</a></div>
    <div class="geo">GEO:
      <span class="latitude">-22.23318</span>,
      <span class="longitude">-49.968899</span>
    </div>
</div>
```

EXEMPLO 4 - MICROFORMATO HCALENDAR – REUNIÃO DO GRUPO DE PESQUISA

Através das figuras 16 e 17, é possível observar o uso do add-on Operator, do Firefox, que recebe e identifica um código do Microformato hCalendar em uma página, e em seguida exporta para algumas ferramentas, como as agendas do Google (gmail) ou Yahoo, além do formato iCalendar (iMac).

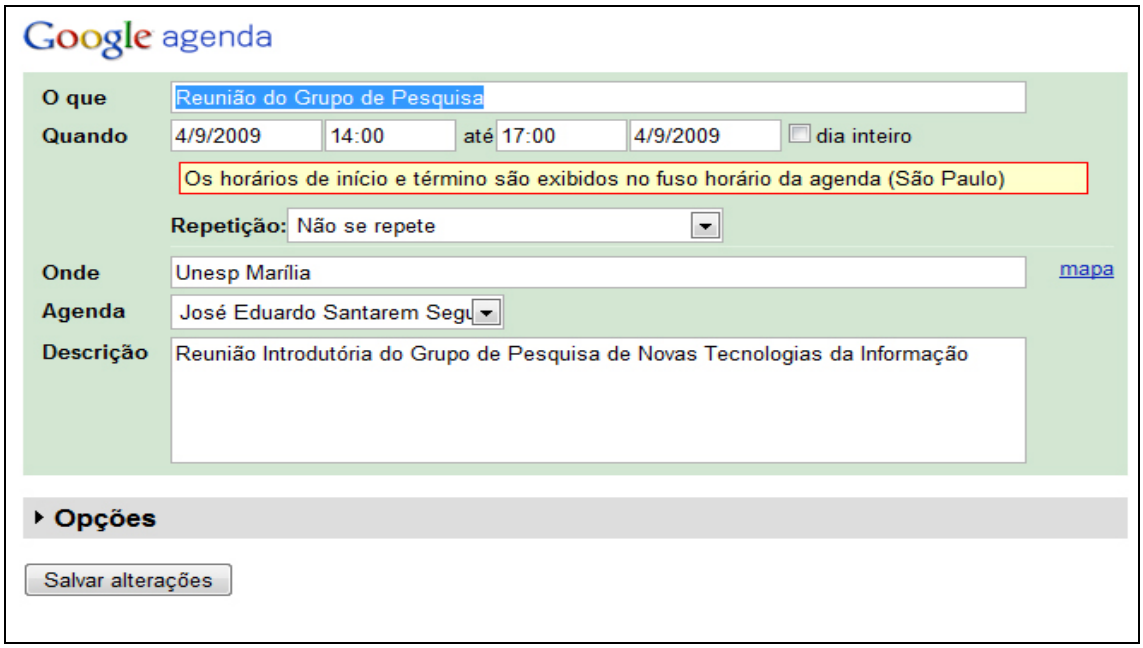


Figura 17 - Agenda do Google recebendo e aguardando usuário salvar a informação do Microformato da reunião.

Fonte: Próprio autor

No código de Microformato, do exemplo 4, utilizado para gerar as imagens 17 e 18, foi agendada uma reunião do Grupo de Pesquisa, e dessa forma, estando o código do Microformato embutido na página HTML, o usuário consegue rapidamente exportar a informação para sua agenda específica. Neste código (exemplo 4) foi incluída a informação geográfica do local da reunião, permitindo ao Operator acessar essa informação e remeter o usuário diretamente ao local determinado no Google Maps (figura 18).

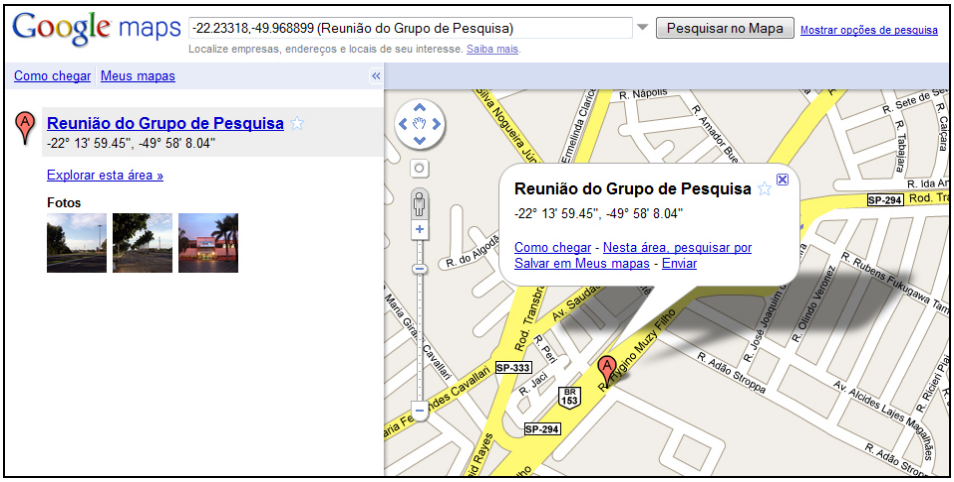


Figura 18 – Google Maps (Mapa localizado através do microformato do exemplo 4).

Fonte: Próprio autor

Além dos Microformatos apresentados, os pesquisadores Eva Méndez, Alejandro Bravo e Leandro Mariano López apresentam, em um artigo denominado “Microformatos: web 2.0 para el Dublin Core”, uma sugestão de microformato para o padrão Dublin Core. Apesar de a especificação ainda não aparecer na lista de drafts do site oficial dos Microformatos, é importante abordar o trabalho desenvolvido, visto que cria um novo conceito para embutir informações sobre um recurso na Web, através do uso de dois padrões consolidados: Dublin Core, para descrição de recursos digitais na Web, e Microformatos, para embutir metainformações em ambientes web.

Mendez, Bravo e Lópes (2007, p. 110) referem-se à criação dos Microformatos DC:

> Desta forma, nós juntamos todos os elementos de metadados DC na lista de Microformatos, e dessa forma permitir reforçar as indiscutíveis vantagens do padrão DC (simplicidade, flexibilidade e adequação para qualquer domínio) para descrever, através dos Microformats DC, qualquer recurso que se deseja citar em um documento XHTML.

A utilização do Microformato DC também já apresenta ferramentas para auxiliar na criação e uso da especificação. A ferramenta para gerar um Microformato DC é o “Dublin Core Metadata Gen: Generator of metadata using Dublin Core” (http://webposible.com/utilidades/dublincore-metadata-gen/index.php?lang=en). Através dos add-nos “Flock” e “Dublin Core Viewer Extension” é possível identificar os recursos descriminados através do Microformato DC na utilização do browser Firefox, como é possível verificar, na figura 19, um pequeno símbolo na canto inferior direito, para que através dele seja apresentado o quadro com as informações do recurso.

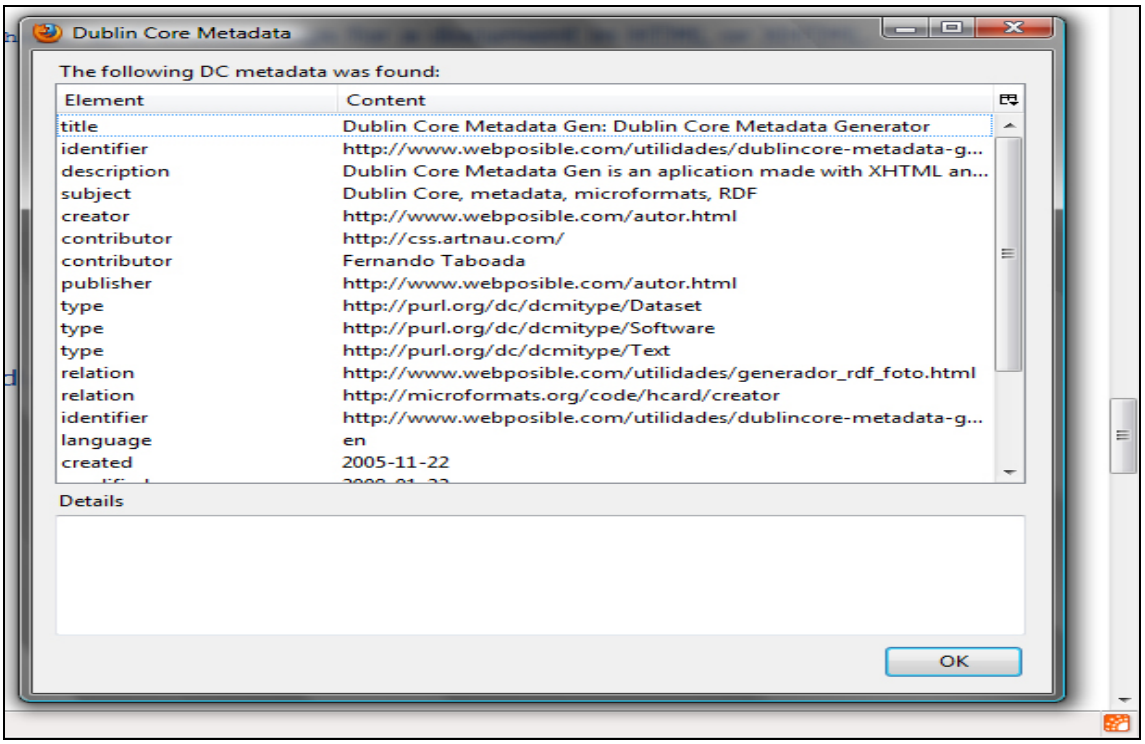


Figura 19 – Dublin Core Viewer Extension.

Fonte: Próprio autor

Os Microformatos se apresentam como uma aplicação real baseada nos conceitos da Web 3.0 e Web 2.0, tornando o conteúdo dos documentos disponíveis na Web mais estruturado e com mais informações.

Considera-se que o uso de Microformatos pode enriquecer muito um ambiente informacional digital como repositórios, e por isso se aborda o assunto como um recurso importante como aplicação prática de parte do modelo de Web Semântica proposto por Berners-Lee.

Se a Web 3.0 pode ser minimamente apresentada através dos microformatos, não será através deles que se obterão resultados de recuperação semântica, visto que o objetivo principal é a separação da estrutura e do conteúdo de um ambiente Web. É imprescindível o uso de ontologias para que um ambiente Web possa utilizar efetivamente a recuperação semântica da informação.

As ontologias, que são parte do modelo concebido por Tim Berners-Lee para a construção da Web Semântica, e de muita importância dentro do contexto desta pesquisa, serão abordadas no próximo capítulo.

4 ONTOLOGIAS: CONCEITOS, LINGUAGENS E FERRAMENTAS

Nos capítulos anteriores verificou-se que grande parte dos esforços de pesquisa relacionados a informações disponíveis na web estão concentrados justamente na construção de um ambiente estruturado de informação, com objetivo de proporcionar uma melhor recuperação da informação.

Neste capítulo será apresentada uma introdução teórica sobre ontologias, visto que esse conceito é ponto-chave na construção de um ambiente informacional digital semântico. Por meio de um levantamento bibliográfico, serão abordados os conceitos mais empregados para o tema e apresentadas as linguagens e as ferramentas mais utilizadas para a construção de ontologias. Também se enfatizará a linguagem OWL, indicada pelo W3C como principal e mais completa linguagem para construção de ontologias.

A verificação das principais tecnologias e métodos disponíveis supõe que apenas a comparação sintática entre termos não atende a principal demanda de recuperação da informação, que é oferecer, como resultado de uma expressão de busca, os principais documentos que estejam diretamente ligados a essa expressão.

Alguns modelos de recuperação têm se apresentado melhor do que outros em circunstâncias diferentes, porém verifica-se que as comparações sintáticas sempre relacionam termos que têm a mesma grafia, não fazendo relação entre termos que têm relação semântica, que é o processo utilizado pelo cérebro para distinguir relações de proximidade entre objetos de um modo geral.

Utilizar ontologias e suas relações é uma das maneiras de se construir uma relação entre termos dentro de um domínio, visto que elas possibilitam contextualizar dados, tornando mais eficiente a interpretação de documentos pelas ferramentas de recuperação da informação.

A palavra ontologia é encontrada em diversos estudos e ciências. Em virtude de sua recente introdução dentro do contexto da Ciência da Informação, registra-se uma grande quantidade de definições e conceitos.

4.1 Definição de ontologia

O termo ontologia deriva do idioma grego, onto (ser) + logia (estudo), e foi inicialmente difundido dentro dos estudos da Filosofia, para estudar as teorias da natureza da existência.

No dicionário Aurélio, a palavra ontologia está definida como a parte da filosofia que trata do ser enquanto ser, isto é, do ser concebido que tem uma natureza comum, inerente a todos e a cada um dos seres. Em epistemologia, refere-se ao conhecimento e à sabedoria.

Estudos baseados em ontologias têm surgido constantemente nas pesquisas relacionadas à Ciência da Informação e também à Ciência da Computação, permeando várias disciplinas e áreas dentro de cada uma das ciências.

Várias são as definições encontradas e que podem se aplicadas ao termo.

Para Guarino (1998, p.7), ontologia é "uma maneira de se conceitualizar de forma explícita e formal os conceitos e restrições relacionados a um domínio de interesse". Numa visão mais tecnológica, o termo refere-se a um artefato de engenharia que, em uma visão simplista, pode ser descrito como uma hierarquia de conceitos relacionados entre si através de uma classificação de parentesco (hipernímia e hipônimo), também chamada de taxonomia.

A definição de Jacob (2003, p.19) aproxima-se muito do conceito de ontologia que mais se aplica à Ciência da Informação quando no contexto da recuperação semântica de informações.

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

Em Ciência da Computação, o estudo de ontologias está ligado à aquisição do conhecimento a partir de dados semiestruturados, aplicando um conjunto de métodos, técnicas ou processos automáticos ou semiautomáticos. Dentro de Ciência da Computação, o termo "ontologia" é originário dos estudos de Inteligência Artificial.

Dados semiestruturados são um tipo de informação nem completamente não-estruturada, nem estritamente tipada, ou seja, é a informação apresentada através de um conjunto de dados que podem estar divididos entre informações armazenadas em banco de dados, que são estruturadas, e também em informações textuais e outros tipos de objetos digitais, que não são estruturadas, e que ficam associados ao conjunto de informações estruturadas e disponíveis para acesso aos usuários.

Ontologias fornecem o conhecimento estruturado e uma infra-estrutura para integrar bases de conhecimentos, independentes da implementação e constituem uma ferramenta poderosa para suportar a especificação e a implementação de sistemas computacionais de qualquer complexidade. Em alguns casos, esse termo é usado apenas como um nome mais rebuscado, denotando o resultado de atividades familiares como modelagem de domínio e análise conceitual. No entanto, em muitos outros casos, as ditas ontologias apresentam algumas peculiaridades como a forte ênfase na necessidade de uma abordagem altamente formal e interdisciplinar, na qual a filosofia e a lingüística desempenham um papel fundamental (GUIZZARDI, 2000).

Gruber (1993, p.2) define ontologias como uma "especificação explícita de uma conceituação". Uma conceituação pode ser representada como um conjunto de objetos, restrições, relacionamentos e entidades que se assumem necessárias em alguma área de aplicação.

A conceituação de Gruber foi modificada por Borst, definindo ontologias como uma "especificação formal de uma conceituação compartilhada" (BORST, 1997).

Como afirmam Chandrasekaran, Josephson e Benjamins (1999), ontologias tratam da organização de objetos, suas propriedades e seus relacionamentos em um determinado domínio de conhecimento. Além disso, disponibilizam termos potencialmente úteis para descrever o conhecimento sobre um domínio específico.

As diferentes apresentações do conceito de ontologia na literatura enriquecem-se mutuamente e ainda sugerem outras. Para Araujo (2003), ontologia é a representação de um vocabulário, frequentemente especializado em algum domínio ou assunto importante. Mais precisamente, não é o vocabulário que qualifica uma ontologia, mas os conceitos que os termos do vocabulário transmitem. Então, transferindo os termos de uma ontologia de uma linguagem para outra, por exemplo, do inglês para o francês, não muda o conceito ontológico.

Para Daum (2002 apud Araujo, 2003), uma ontologia é uma descrição formal dos conceitos e relacionamentos que existem dentro de um domínio, isso significa que uma ontologia se relaciona com um vocabulário específico e com uma linguagem específica.

O uso de Ontologias torna possível definir uma infraestrutura para integrar sistemas inteligentes no nível do conhecimento (NOVELLO, 2002).

A maneira como Novello aborda o uso de Ontologias cria uma relação direta e faz o termo pertencer ao contexto da informação e tecnologia.

O nível do conhecimento é independente do nível de implementação. Ontologias apresentam grandes vantagens como:

- Colaboração: possibilitam o compartilhamento do conhecimento entre os membros interdisciplinares de uma equipe;
- Interoperação: facilitam a integração da informação, especialmente em aplicações distribuídas;

- Informação: podem ser usadas como fonte de consulta e de referência do domínio;
- Modelagem: as ontologias são representadas por blocos estruturados que podem ser reusáveis na modelagem de sistemas no nível de conhecimento.

Novello (2002) afirma ainda:

> [...] as ontologias podem servir como uma ferramenta navegacional de consulta para o usuário, fornecendo informação semântica sobre restrições, conceitos e relacionamentos do domínio, mantendo o conhecimento do domínio compartilhado entre todos os membros de uma equipe e até mesmo entre equipes geograficamente separadas.

Esta definição de Novello indica que as ontologias podem desempenhar um papel fundamental na relação de um ambiente informacional com seus usuários.

As ontologias apresentam-se como um modelo de relacionamento de entidades em um domínio particular do conhecimento. O objetivo principal de sua construção é a necessidade de um vocabulário compartilhado em que um conjunto de informações possam ser trocadas e também reusadas pelos usuários de uma comunidade. Considere os usuários de uma comunidade seres humanos ou agentes inteligentes.

Guarino (1998, p.10) propõe uma diferenciação entre as ontologias, de acordo com sua generalidade:

- Ontologias de topo ou de senso comum: descrevem conceitos bastante gerais, como espaço, tempo, matéria, objeto, evento, ação, etc., que são independentes de um problema ou domínio particular.
- Ontologias de domínio: descrevem o vocabulário relacionado a um domínio particular, especializando conceitos introduzidos nas ontologias de topo. Exemplos comuns são ontologias de medicina, automobilismo, computação, entre outras.

- Ontologias de tarefa: descrevem tarefas de um domínio, como processos, planos, metas e escalonamentos através de uma visão funcional.
- Ontologias de aplicação: descrevem conceitos que dependem de um domínio e de uma tarefa particular, portanto, geralmente são uma especialização de ontologias de domínio e tarefa. Esses conceitos frequentemente correspondem aos papéis desempenhados por entidades do domínio enquanto executam certa atividade.

Guarino diz ainda que ontologias de domínio e de tarefa especializam os termos presentes nas ontologias de topo, e que, por sua vez, ontologias de aplicação utilizam termos e regras das ontologias de domínio e de tarefas.

A divisão apontada por Guarino deverá ser mais claramente observada quando na construção de ontologia e, principalmente, na utilização de tecnologias que possibilitam a criação real de uma ontologia para um determinado domínio.

Outra característica importante do termo é ressaltada em Freitas (2008), e remete ao fato de que uma ontologia não pode ser tratada apenas como uma hierarquia de conceitos, mas também como um conjunto de relações, restrições, axiomas, instâncias e vocabulário.

Apesar de serem aplicadas em diversas áreas dentro da Ciência da Informação, as ontologias têm um papel especialmente importante para a Web Semântica. De acordo com Berners-Lee, Hendler e Lassila (2001), para o funcionamento da Web Semântica, computadores devem ter acesso a coleções estruturadas de informação e conjuntos de regras que possam usar para conduzir raciocínio automático, sendo esse o principal desafio da área.

Muitos ainda são os conceitos e definições encontrados na literatura sobre ontologia.

Ressalte-se, porém, que apesar dos diferentes vocabulários e vertentes, praticamente todas as definições citam a construção de uma estrutura de relação entre conceitos dentro de um domínio.

A abordagem que se faz em relação à Ontologia é de que essa estrutura de informação está inserida dentro de um contexto de Estruturas de Representação do Conhecimento.

4.2 Estruturas de Representação do Conhecimento

Este trabalho utiliza o termo "Estruturas de Representação do Conhecimento" como forma de unificar nesse conceito estruturas de representação como taxonomias, ontologias e tesauros.

O Enancib, principal evento dos programas de pós-graduação em Ciência da Informação no Brasil tem destacado alguns temas emergentes como taxonomias, ontologias e Web Semântica nas publicações a respeito do termo "Representação do Conhecimento", nos anos de 2005, 2006 e 2007. (FUJITA, 2008).

4.2.1 Vocabulário Controlado

O vocabulário controlado é um instrumento terminológico para definir os termos e limites de um determinado domínio de conhecimento.

Segundo definição da organização norte-americana National Information Standards Organization, presente no documento que propõe as linhas gerais para a construção, formatação e manutenção de vocabulários controlados monolingües (ANSI/NISO Z39-19-2005), um vocabulário controlado é uma lista finita de termos que tem seus respectivos significados explicitados com o intuito de evitar redundâncias e ambigüidades, utilizados para representar informações de maneira padronizada (RAMALHO, 2006).

Segundo Kobashi (2008, p.1), vocabulário controlado é

> [...] uma LINGUAGEM ARTIFICIAL constituída de termos organizados em estrutura relacional. Um vocabulário controlado é elaborado para padronizar e facilitar a entrada e a saída de dados em um sistema de informações. Tais atributos promovem maior precisão e eficácia na comunicação entre os usuários e o sistema de informações.

Exitem ainda outras definições para vocabulário controlado, como a apresentada por Lima e Boccato (2009, p. 133):

> O vocabulário controlado, como toda linguagem documentária, é um instrumento de organização e recuperação da informação, construído com a finalidade de propiciar a representação e a recuperação dos conteúdos informacionais dos documentos cadastrados.

Através das afirmações apresentadas, verifica-se que os vocabulários controlados são instrumentos que condicionam e permitem a padronização de um sistema de informação.

Os vocabulários controlados são estruturados para possibilitar diferentes tipos de relacionamentos entre termos, determinando desde níveis de relacionamentos simples até estruturas mais complexas (ANSI/NISO Z39-19-2005).

Um vocabulário controlado é composto de termos que são organizados de forma hierárquica, afirma Kobashi (2008, p.1):

> Todo vocabulário controlado é composto por um conjunto de termos que representam conceitos de um ou vários campos de conhecimento. Tais signos são dispostos em estrutura relacional previamente definida. Em geral, os vocabulários controlados são apresentados em ordem hierárquica e alfabética (macroestrutura e microestrutura).

Os vocabulários controlados, apesar de utilizados em ambientes mais restritos, podem ser aplicados na construção de qualquer tipo de base de conhecimento.

Kobashi (2008, p.1) indica as funções de um vocabulário controlado dentro de um ambiente informacional.

> Uma das funções do vocabulário é REPRESENTAR a informação e o conhecimento por meio de um conjunto controlado e finito de termos – os descritores.
>
> CONTROLAR ou padronizar é outra função básica de um vocabulário controlado. A localização ou identificação de informação, sem padronização léxica, torna-se errática. Resultados eficientes de busca dependem, assim, de coincidência entre as formas de representação utilizadas pelo sistema de informação e pelo usuário. Um vocabulário controlado, portanto, garante a comunicação efetiva entre sistema de informação e usuário.

Tálamo, Lara e Kobashi (1992, p. 1999) afirmam:

> [...] cabe a terminologia, desse modo, operar ao nível sintático-semântico, produzindo terminologias específicas de acordo com o estado-da-arte de cada campo considerado. Tais repertórios ou listas de termos especializados de um domínio particular são acomanhados de definições que remetem o termo ao seu referente [...]

A abordagem em relação aos vocabulários controlados dentro dessa pesquisa se dá pela necessidade da construção de um ambiente informacional digital que considere as relações semânticas entre termos.

É importante ressaltar que a construção estruturas de representação do conhecimento é trabalho de profissionais especializados, que conhecem primeiramente as características de construção desse tipo de instrumento, e de profissionais que tenham claramente definida a estrutura informacional do domínio a qual será aplicado o vocabulário controlado.

Kobashi (2008, p.2) faz referência à construção de vocabulários controlados.

> Para ser útil, deve refletir, de um lado, os objetivos do sistema de informação para o qual foi elaborado e, de outro, a linguagem dos usuários. Por essa razão, sua construção é coletiva, requer trabalho integrado, colaborativo, envolvendo tanto os gerenciadores do sistema de informação quanto os usuários da informação. Além disso, é uma linguagem dinâmica que se desenvolve em consonância com a dinâmica das áreas de conhecimento representadas no sistema de informação. Necessita, portanto, de atualização periódica

Os vocabulários controlados, assim como as ontologias são tipos de estruturas de representação do conhecimento, assim como também são os tesauros e as taxonomias, pois todos têm o objetivo de apresentar, relacionar e controlar as informações dentro de um domínio do conhecimento. Portanto, far-se-á uma abordagem a respeito desses outros instrumentos que também têm objetivos semelhantes aos das ontologias.

4.2.2 Tesauro

Os tesauros são uma espécie de linguagem especializada que foi apresentada pela primeira vez há quase dois séculos.

A palavra tesauro tem origem do latim thesaurus, que significa tesouro. Foi empregada como título no dicionário analógico de Peter Mark Roget, "Thesaurus of English words and phrases", publicado em Londres, pela primeira vez, em 1852. O autor era secretário da Royal Society e objetivava facilitar sua atividade literária. Trabalhou nesse projeto durante 50 anos. Em seu dicionário, as palavras foram agrupadas em ordem distinta da alfabética. Priorizaram-se as ideias que exprimiam e esta foi a ordem escolhida. A busca por palavras dava-se sempre por aquilo que elas podiam expressar, com seu significado (GOMES, 1990).

Segundo Ramalho (2006, p.91),

> Quanto aos tesauros a norma ANSI/NISO Z39-19-2005 define como um vocabulário controlado organizado segundo uma ordem conhecida e estruturada com o intuito de disponibilizar claramente os relacionamentos de equivalência, associação, hierárquicos e homônimos existentes entre termos. Os tesauros também comportam características de taxonomias com um conjunto de relacionamentos semânticos, visando garantir que os conceitos e seus relacionamentos sejam descritos de maneira consistente em um sistema de classificação e recuperação de informações.

A principal característica de um tesauro está na construção de uma estrutura que relacione e defina termos dentro de um domínio do

conhecimento, de forma que as associações entre os termos utilize uma estrutura relacional hierárquica e associativa de informações.

De acordo com Sales e Café (2009, p. 102),

> Tesauros são vocabulários controlados formados por termos-descritores semanticamente relacionados, e atuam como instrumentos de controle terminológico. Os tesauros podem estar estruturados hierarquicamente (gênero-espécie e todo-parte) e associativamente (aproximação semântica), e são utilizados principalmente para indexar e recuperar informações por meio de seu conteúdo.

Neste contexto os tesauros caracterizam-se por relações hieráquicas (herança) e também semânticas.

Segundo Moreira e Moura (2006, p.2),

> Um tesauro é uma linguagem de documentação com a característica específica de possuir relações entre os termos que o compõem. O termo linguagem de documentação compreende, genericamente, os sistemas de classificação bibliográfica, as listas de cabeçalho de assunto e os tesauros, os quais surgiram estimulados pela necessidade de manipulação de grande quantidade de documentos de conteúdos especializados. Os tesauros constituem uma ferramenta de indexação já consolidada nas atividades de organização da informação empregada por muitos que exercem essas atividades.

Conclui-se, portanto, que os tesauros atuam na linguagem de indexação de documentos.

A utilização de tesauros fortalece a base de conhecimento na qual os documentos são depositados e seu uso tende a facilitar a descrição e, consequentemente, a recuperação da informação. Um tesauro bem construído e que consiga relacionar os principais termos de um domínio de conhecimento facilita o acesso à informação. Porém um tesauro que não atende aos requisitos mínimos de envolvimento no domínio a que está proposto, ou que não recebe uma atualização devida, de acordo com a atualização constante das áreas, pode representar justamente o inverso, ou seja, uma estrutura “dura” de descrição da informação e que em muitos casos, além de não atender ao contexto de conhecimento, também dificulta o

processo de recuperação da informação, visto que os termos não indexam devidamente os documentos.

Tálamo, Lara e Kobashi (2002, p. 198) apontam para isso,

> Na prática, o uso do tesauro fica comprometido pelo aparecimento de qualidades do texto individual que não são passíveis de serem enquadrados em parâmetros prévios e preditivos. Assim, no lugar de uma análise da significação discursiva com referência às circunstâncias de emissão, supõe-se uma interpretação amarrada em definições conceituais (das propriedades da palavra) quase sempre obscuras ou intuídas, já que, muitas vezes, as relações semânticas entre os descritores não são suficientemente claras e rigorosas. Perde-se, desse modo, a informação específica e individual do texto, em prol de uma atribuição de sentido prevista e sedimentada fora das circunstâncias de enunciação.

Dessa forma, fica claro que a construção e atualização de um tesauro implica diretamente nos resultados obtidos através da construç ão da informação baseada nessa linguagem de indexação.

Apesar de dividir objetivos semelhantes com as ontologias, são linguagens diferentes, que buscam evoluir no processo de descrição e recuperação da informação, sempre em busca de minimizar a discrepância entre a necessidade do usuário e o resultado de suas buscas.

Sales e Café (2009, p.101) abordam este contexto da seguinte maneira:

> O tesauro é uma linguagem documentária caracterizada pela especificidade e pela complexidade existente no relacionamento entre os termos que comunicam o conhecimento especializado. A ontologia é um modelo de representação do conhecimento que, a exemplo do tesauro, é utilizada para representar e recuperar informação por meio de estruturas conceituais (no caso da ontologia o meio de ação é o informático).

Moreira, Alvarenga e Oliveira (2004, p.21) também analisam as semelhanças e diferenças entre ontologias e tesauros.

> [...] a análise quantitativa evidenciou a diferença de propósitos entre os dois instrumentos. A freqüência de ocorrência de termos, bem como a abrangência das definições sobre as categorias, mostrou que os tesauros têm como propósito,

> servir de instrumento de registro terminológico e para ser usado por pessoas, e não para registro do conhecimento para inferências computacionais. Por exemplo, nas definições sobre o termo "tesauros", a ocorrência de termos como 'usuário' e "usuários", é uma evidência no sentido do uso com sistemas de classificação e recuperação de documentos. Já no caso das definições sobre ontologia, a ocorrência de termos tais como "formal" e "Lógica", demonstra a necessidade de registro do conhecimento do domínio em uma linguagem que possa ser processada pelo computador para a realização de inferências.
>
> No entanto, semelhanças também foram encontradas, uma ontologia como vista pela ciência da computação é um sistema de conceitos, da mesma forma que os tesauros, e como tal pertence ao nível epistemológico e não ao ontológico. A diferença em relação aos tesauros pode ocorrer em termos de linguagem, de nível de formalização e de propósitos. Neste sentido pode ser adequado que, no âmbito da ciência da computação, os tesauros sejam enquadrados como ontologias.

Portanto, ressalte-se que ontologias e tesauros são estruturas diferenciadas com objetivos semelhantes, sendo que as ontologias são, em vários momentos, encontradas como recursos informáticos para representação do conhecimento.

4.2.3 Taxonomias

As taxonomias são compostas de termos ou conceitos sobre o universo da informação armazenada, relacionados de forma hierárquica. O termo taxonomia foi inicialmente utilizado para definir uma estrutura hierárquica que separava os seres vivos de acordo com suas características em comum.

Segundo Campos e Gomes (2008, p.1),

> Taxonomia é, por definição, classificação, sistemática e está sendo conceituadas no âmbito da Ciência da Informação como ferramenta de organização intelectual. É empregada em portais institucionais e bibliotecas digitais como um novo mecanismo de consulta, ao lado de ferramentas de busca.

Segundo a norma ANSI/NISO Z39-19-2005 (p.9), uma taxonomia é, "Uma coleção de termos de um vocabulário controlado organizada em uma estrutura hierárquica".

Ramalho (2006, p.91) afirma que,

> As taxonomias permitem classificar informações em uma estrutura de árvore, por meio de relacionamentos de generalização ("pai-filho", "tipo-de"), não possibilitando atribuir características ou propriedades aos termos nem expressar outros tipos de relacionamentos.

Taxonomia é uma forma de classificar ou categorizar um conjunto de coisas em uma hierarquia. Tem a mesma estrutura de uma árvore, constituída por uma raiz e ramificações, onde cada ponto (cada nó) é uma entidade de informação. No contexto das tecnologias da informação, uma taxonomia é geralmente entendida como a classificação das informações ou entidades, sob a forma de uma hierarquia, de acordo com a presumível relação de entidades do mundo real que elas representam (tradução nossa) (DACONTA, et al., 2003).

Dessa forma, vê-se a taxonomia como um mecanismo de sistematizar informações através de categorias, ou seja, como um modelo de classificação hierárquica que possibilita a identificação, localização e estudo dos dados.

Dentro do contexto da Ciência da Informação, as taxonomias atualmente são estruturas classificatórias que têm por finalidade servir de instrumento para a organização e recuperação de informação em empresas e instituições. Estão sendo vistas como meios de acesso, atuando como mapas conceituais dos tópicos explorados em um serviço de recuperação. O desenvolvimento de taxonomias para o tipo de negócio da empresa tem sido um dos pilares da gestão da informação e do conhecimento. (BAYLEY 2007)

Entende-se que as ontologias podem ser também uma forma de representação e aplicação computacional das taxonomias.

4.3 Composição e Construção de Ontologias

O objetivo da criação de uma ontologia é dividir o conhecimento de um domínio de interesse comum e prover um entendimento unificado de definições de termos de um domínio, além de especificar relações entre estes termos.

A construção de uma ontologia pode ser pensada como um conjunto de peças que formam uma estrutura completa. Assim, ela pode ser separada e apresentada como um conjunto de componentes.

Os componentes básicos de uma ontologia são: classes/conceitos (organizadas em uma taxonomia), relações (representam o tipo de interação entre os conceitos de um domínio), axiomas (usados para modelar sentenças sempre verdadeiras) e instâncias/indivíduos (utilizadas para representar elementos específicos, ou seja, os próprios dados). (GRUBER, 1996).

A abrangência da ontologia é definida como domínio. Domínio é a expressão que define uma parte de um ambiente ou do mundo, onde se estabelecem claramente os limites, ou seja, onde é possível definir exatamente o conjunto de informações que se pretende tratar.

As classes e instâncias compõem o vocabulário. Classes são sinônimos de categorias. As classes definem os conceitos dentro do domínio considerado, e também podem ser interpretadas como uma estrutura modular completa, que descreve as propriedades estáticas e dinâmicas dos elementos em um domínio. Uma classe abstrai um conjunto de objetos com características similares.

Toda classe é caracterizada por seus atributos, que podem ser chamados também de propriedades de uma classe. São os atributos que dão características diferentes a cada classe. Quando uma classe é instanciada, cada um dos atributos recebe valores.

É possível estabelecer relações de hierarquia entre as classes e são essas relações que formam a taxonomia dentro de um domínio. Neste

conceito de relação hierárquica de classes, denominado herança, as classes estabelecem relações que são chamadas de pais e filhos, e as classes filho herdam as características, atributos, das classes pai.

As instâncias são as ocorrências particulares do objeto em relação à classe considerada, chamadas de indivíduos. Uma instância pode ser definida como a materialização de uma classe. Uma instância também descreve conceitos, mas de forma individualizada, única e concreta, fazendo referência a um objeto real. Numa descrição abstrata da dualidade classe-instância, a classe é apenas uma matriz estrutural, que especifica objetos, mas que não pode ser utilizada diretamente; a instância representa o objeto concretizado a partir de uma classe, que pode ser vista como um protótipo.

Para permitir o enriquecimento semântico de uma ontologia são estabelecidas regras, que impõem restrições as suas classes e atributos, ou seja, são assertivas lógicas que estabelecem limites e obrigam ou permitem valores tanto para a classe como para os atributos.

Uma ontologia é uma estrutura de classe para representar uma realidade através de uma linguagem formal, composta de vocabulário (classes e instancias), relacionamentos (herança e relação entre as classes, que são as taxonomias) e regras (limites estabelecidos para classes e atributos).

Dada a estrutura de composição das ontologias, como já visto, elas se apresentam como um modelo de relacionamento de entidades e suas interações, em algum domínio particular do conhecimento ou específico a alguma atividade. O objetivo da construção de ontologias está diretamente ligado à necessidade de um vocabulário compartilhado para se trocarem informações entre os membros de uma comunidade, sejam eles humanos ou agentes inteligentes.

Neste caso, entende-se a taxonomia como um modelo conceitual, e as ontologias como formas tecnicamente aplicáveis destes modelos, porém em formatos que podem ser utilizados em ambientes digitais, como, por

exemplo, os repositórios digitais, além de outras estruturas e ambientes de informação.

São inúmeros os benefícios quando se define um domínio de interesse com ontologias: compartilhamento do conhecimento, aplicação de uma ontologia genérica para um domínio de conhecimento específico e compreensão semântica dos dados do domínio. Para garantir que uma ontologia seja construída com qualidade, é necessário definir o domínio de conhecimento com objetividade, descrevendo o conhecimento essencial ao domínio e definindo um vocabulário que evite interpretações ambíguas (GRUBER, 1993).

Se os benefícios forem claros, o mesmo não pode se dizer sobre a sua construção. Grande parte da dificuldade do desenvolvimento de ontologias paira sobre sua construção. Essa dificuldade para se construir ontologias fica evidente, principalmente porque motiva a demora para o estabelecimento de uma estrutura clara e de fácil utilização da Web Semântica.

Alguns trabalhos propõem metodologias diferentes para a construção de ontologias, e, mesmo assim, ainda não se tem uma definição sobre a melhor maneira de construí-las, ou seja, não existe a melhor forma.

Outra dificuldade encontrada na criação de ontologias é que grande parte das propostas de metodologias prevêem a construção manual, com auxílio de algumas ferramentas, porém a construção manual de ontologias é um processo complexo, tedioso e de alto custo, e, por ser extremamente artesanal, também propensa a erros. (BREWSTER; CIRAVEGNA; WILKS, 2003).

Diversos trabalhos vêm propondo a construção automática ou semiautomática de ontologias, para agilizar o processo e auxiliar na atualização das mesmas.

Este capítulo abordará mais amplamente a construção de ontologias de maneira manual, visto que a criação automática de ontologias

parece ser um processo mais demorado para se estabelecer ante o processo manual.

Como o processo de construção de ontologias ainda não está totalmente estabelecido, é possível encontrar desenvolvedores utilizando seus próprios critérios e métodos para o desenvolvimento.

É importante ressaltar que a construção de ontologias deve estar sempre condicionada à aquisição do conhecimento sobre o domínio estabelecido e, posteriormente, à implementação da estrutura de classes que vai compor a ontologia.

O processo de construção de ontologias está diretamente ligado ou condicionado à utilização de linguagens de marcação semântica que foram construídas com esse propósito, ou seja, que suportem estruturas para representação do conhecimento. As linguagens utilizadas devem permitir descrição formal de um conjunto de termos sobre um domínio específico, ser compatíveis com a Web, ter uma sintaxe e uma semântica bem definida e, principalmente, suportar raciocínio eficiente.

4.4 Linguagens de Marcação Semântica

As linguagens de marcação semântica tiveram início com a criação da linguagem KIF (Knowledge Interchange Format), que teve propósito inicial nos princípios da inteligência artificial e foi desenvolvida em 1992. A linguagem KIF pouco serviu para desenvolver ontologias, porque o processo de criação utilizando a linguagem era muito complexo e trabalhoso, porém serviu como base para criação da Ontolíngua, que foi desenvolvida como uma camada sobre a linguagem KIF.

As linguagens de marcação semântica para a construção de ontologias para web devem garantir distinção entre as classes, propriedades e relações, de modo a evitar ambiguidades durante o desenvolvimento.

A primeira linguagem a se destacar com o objetivo de descrever recursos da Web foi a RDF (Resource Description Framework), desenvolvida pelo W3C e recomendada pelo mesmo consórcio, no ano 2000. Conhecida pela falta de expressividade em suas representações, a linguagem RDF recebeu como complemento o RDF-Schema, que dá à linguagem RDF o poder de construção de estruturas como hierarquias, propriedades e subpropriedades, entre outros, que a linguagem RDF até então no possibilitava.

O uso conjunto da linguagem RDF + RDF Schema é denominado RDFS e serviu como base para o desenvolvimento de outras linguagens e soluções para construção de ontologias, cada uma delas com suas vantagens e facilidades, como: SHOE (Simple HTML Ontology Extensions), que foi a primeira linguagem de ontologia desenvolvida especificamente para Web Semântica; OIL (Ontology Inference Layer), que foi desenvolvida através de um esforço conjunto de universidades da Europa; XOL (Ontology Exchange Language), DAML (DARPA Agent Markup Languagem), desenvolvida pela americana DARPA; DAML e OIL (DAML+OIL), que, combinadas, também formaram uma nova linguagem, e, desde fevereiro de 2004, recomendada pelo W3C. A linguagem que mais vem sendo utilizada para construção de ontologias é a OWL (Web Ontology Language).

4.4.1 RDF e RDF Schema

Segundo o W3C, o RDF é uma linguagem de uso geral para representar informações na Web. O RDF tem como princípio fornecer interoperabilidade aos dados, de forma que possa contribuir com a recuperação de informações de recursos na Web.

Segundo Lassila (1999),

> RDF é uma aplicação da linguagem XML que se propõe ser uma base para o processamento de metadados na Web. Sua padronização estabelece um modelo de dados e sintaxe para codificar, representar e transmitir metadados, com o objetivo

> de torná-los processáveis por máquina, promovendo a integração dos sistemas de informação disponíveis na Web. (tradução nossa)

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

O modelo RDF é constituído de três objetos básicos: recursos, propriedades e declarações. Um recurso é uma informação (página web, livro, cd, pessoa, lugar, documento disponível em um repositório ou biblioteca digital) que pode ser identificada por uma URI (Universal Resource Identificator). Propriedades são as informações que representam as características do recurso, ou seja, são os atributos que permitem distinguir um recurso de outro ou que descrevem o relacionamento entre recursos. A declaração é a constituição da informação completa, que compreende um recurso com suas propriedades e valores para as propriedades.

Uma URI pode ser um local ou página na WEB como uma URL (Unified Resource Locator) ou ainda outro tipo de identificador único.

Os três objetos citados — recurso, propriedade e declaração — são normalmente referenciados também como sujeito, predicado e objeto, formando o modelo básico primitivo do RDF, que é constituído de registros com objeto, propriedade e valor. Basicamente, a representação de uma sentença em RDF é feita utilizando-se um grafo. Um grafo é um modelo matemático muito poderoso que pode ser aplicado na resolução de um conjunto de problemas. É composto por um conjunto de vértices e arestas/arcos.

Além de representar graficamente uma informação através de grafos, o modelo RDF pode ser representado através da sintaxe XML. O modelo de representação de RDF através da linguagem XML demonstra que o RDF é uma linguagem muito mais indicada para representação de metadados do que propriamente para linguagem de ontologias.

Lassila (1999) relata que

> a especificação do W3C apresenta duas sintaxes de XML para codificação de um modelo de instância de dados em RDF: a sintaxe de serialização e a sintaxe abreviada. A diferença mais marcante entre as duas está em como a estrutura do modelo RDF é apresentada. A primeira nos oferece uma estrutura mais completa enquanto a segunda nos oferece uma forma mais compacta.

A seguir, uma representação gráfica e com linguagem XML para uma sentença, apresentada por Santarem Segundo (2004).

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

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

Esta sentença pode ser representada pelo diagrama da figura 20:

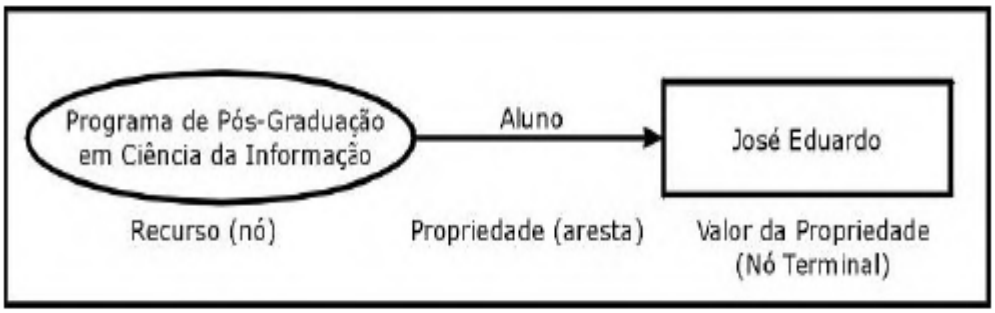


Figura 20 – Diagrama RDF

Fonte: Próprio autor

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

A sentença pode ser também apresentada através da linguagem XML, como no exemplo 5:

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<rdf:RDF xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:about="Programa de Pós-Graduação em CI">
        <f:aluno>
                José Eduardo
        </f:aluno>
</rdf:Description>
</rdf:RDF>
```

EXEMPLO 5 – SENTENÇA RDF

Como foi visto, a linguagem RDF fornece um limitado número de elementos predefinidos, inviabilizando o desenvolvimento de vocabulários próprios por comunidades independentes, e não apresenta subsídios necessários para constituição de uma linguagem de ontologias, sendo sugerida a extensão da linguagem.

Segundo Silva e Lima (2002, p.2),

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

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

A RDF Schema é responsável por prover mecanismos para declaração dessas propriedades. Um esquema não define somente as propriedades dos recursos, mas também os tipos de recursos que estão sendo descritos. Pode ser entendido como uma espécie de dicionário onde são definidos os termos que serão utilizados em declarações RDF. A especificação da RDF Schema do W3C fornece os mecanismos necessários à definição de elementos, de classes de recursos, de possíveis restrições de classes e relacionamentos e detecção de violação de restrições (BRICKLEY e GUHA, 2000).

O RDF-Schema tem disponível um framework que permite descrever as classe e as propriedades, ampliando a gama de informações que podem ser descritas através da linguagem RDF.

Segundo Harman e Koohang (2007, p. 300),

> Usando RDF Schema, a semântica e as propriedades de ambos os elementos de um vocabulário podem ser expressos através de um framework único. RDF Schema permite a descrição das relações entre os termos não só dentro de um único padrão, mas em cruzamento com outros padrões. Também permite a descrição de qualquer número de atributos do vocabulário, termos próprios, utilizando as propriedades RDF.

Os autores afirmam ainda que “RDF Schema possui a base semântica que é utilizada em praticamente todas as descrições realizadas em RDF, englobando tanto as propriedades mais refinadas e as subclasses.”

Como já foi citado, o conjunto RDF e RDF Schema, quando utilizados em conjunto, são denominados RDFS.

Apesar de todas as possibilidades criadas com a criação do RDF Schema, que estende as características de uso da linguagem RDF, o RDFS ainda é considerado limitado para a criação de ontologias, pela falta de conectivos lógicos, falta de expressividade de seus construtores, restrições de existência ou cardinalidade e falta de propriedades transitivas, inversas ou simétricas.

Na construção da estrutura da Web Semântica, essa falta de recurso da RDFS fica ainda mais clara, pois acima da camada destinada ao RDF fica uma camada de ontologia, separando a camada de esquema da camada lógica, demonstrando que, sozinha, a linguagem RDF não pode descrever ontologias.

4.4.2 Simple HTML Ontology Extensions (SHOE)

A linguagem SHOE, projeto da Universidade de Maryland, é uma extensão da linguagem HMTL, que tem como princípio inserir no código HTML informações que possam representar ontologias. Essas informações são inseridas através de tags próprias que não são conhecidas da linguagem HTML, e não são interpretadas pelo browser, servindo neste caso como marcações semânticas que poderão ser interpretadas por máquinas ou outros tipos de recuperação de informações disponíveis na Web, que não os sintáticos propriamente ditos. A linguagem, depois de criada, recebeu uma adaptação para permitir compatibilidade com XML.

O funcionamento da linguagem é baseado em um mecanismo de definição de ontologias, instâncias de ontologias e instâncias de dados em páginas Web. Para definir sua estrutura, a linguagem SHOE faz distinção entre os conteúdos das páginas (asserções ou instâncias) e a terminologia (informações sobre metadados).

A linguagem SHOE apresenta uma grande dificuldade de manutenção, e esse foi um dos motivos que levou o projeto a ser descontinuado, migrando os pesquisadores para os estudos sobre DAML+OIL e OWL.

4.4.3 Ontology Inference Layer (OIL)

A falta de semântica da linguagem RDF, que impedia o suporte a mecanismos de inferência, foi uma das principais causas do desenvolvimento da linguagem OIL.

A linguagem OIL foi desenvolvida para ser compatível com os padrões do W3C, incluindo RDF e XML, e para explorar as primitivas de modelagem da linguagem RDFSchema. Isso indica que toda aplicação que suporta RDF pode entender, pelo menos minimamente, um documento OIL.

A linguagem OIL foi inicialmente desenvolvida com apoio e patrocínio de um consórcio da comunidade européia, fazendo parte do projeto On-to-knowledge, e teve como principal requisito a facilidade de adoção por parte dos desenvolvedores, servindo principalmente à comunidade ligada à Web semântica (HORROCKS et al 2001). Os principais integrantes do projeto OIL são: a University of Manchester (Inglaterra), Vrije Universiteit Amsterdam (Holanda), Stanford University (EUA), University of Karlsruhe (Alemanha), Administrator Nederland (Holanda), Research Bell Labs (EUA) e o MIT (EUA).

As principais características do projeto, além das já descritas anteriormente, são:

- Lógica de descrição, suportando dessa forma inferência e fornecendo semântica formal;
- Permitir definições baseadas em frames, fornecendo primitivas de modelagem epistemológica e também definições em lógica de descrições.
- As definições de ontologias são geradas sobre XML e RDF.
- Inferência: apesar de perder um pouco de expressividade em relação à Ontolingua, tem, em contrapartida, um motor de inferência consistente, completo e eficiente, capaz de manipular tanto frames quanto lógica de descrições.

OIL foi projetada para ser um padrão extensível. Para tanto, OIL é estruturada em camadas:

- O nível mais baixo, chamado Core OIL, é compatível com RDF Schema. Ontologias definidas pelo Core OIL podem ser interpretadas por qualquer aplicação que dê suporte à RDF Schema.
- O próximo nível, denominado Standard OIL, adiciona funcionalidades, tornando OIL apenas parcialmente compatível com RDF Schema. Essa camada é desenvolvida para prover expressividade e formalismo suficiente para permitir raciocínio e dedução.

Uma ontologia escrita em OIL é constituída de três partes principais:

- O *container* ou recipiente, que provê a estrutura de metadados da ontologia, ou seja, como ela deverá ser apresentada. Neste caso, o OIL utiliza o padrão Dublin Core para definição dos metadados. Os metadados apresentados serão título, autor, assunto, etc.
- A *definition* ou definição da ontologia consiste na definição efetiva dos conceitos da ontologia. Essa definição deverá ser apresentada através de um conjunto de expressões que descrevem classes e slots. São definidos três tipos expressões: class definition, slot constraints e slot definition.
- A terceira parte é representada por um nível chamado de objeto, que provê o armazenamento de instâncias, porém ele só é implementado através das sublinguagens específicas Heavy Oil e Instance Oil.

Várias ferramentas foram disponibilizadas pela comunidade de pesquisadores da linguagem OIL para edição e verificação de ontologias, entre elas destacam-se: OntoEdit, OILEd e Protegé-2000.

4.4.4 DAML e DAML+OIL

A agência norte americana DARPA, que participou diretamente no início do desenvolvimento da Internet, em conjunto com o consórcio W3C constituíram a linguagem DARPA Agent Markup Language (DAML), que foi baseada nos esforços empregados e na experiência adquirida em tecnologias e linguagens, como: XML, RDF, OIL, SHOE e RDFS.

O objetivo era a construção de um framework unificado para uma linguagem de representação de ontologias para a web, estendendo a linguagem RDF de forma a deixá-la mais expressiva.

A linguagem DAML é muito similar a OIL, principalmente no que diz respeito às funcionalidades.

Entre as principais características similares, verificam-se: oferecimento de propriedades transitivas e inversas, suportam hierarquia de conceitos e propriedades, oferecem apoio a tipos de dados concretos como inteiros e listas.

A primeira especificação de DAML, lançada em Outubro de 2000, foi conhecida como DAML-ONT, e apenas dois meses depois substituída por uma nova versão denominada DAML+OIL. A fusão das linguagens DAML e OIL permitiu a criação de uma linguagem mais consistente e mais clara.

A especificação DAML+OIL continuou recebendo atualizações. Em março de 2001 passou a ser dividida em duas partes: domínio de objetos (object domain), que se baseia em objetos que são membros de classes definidas na ontologia de DAML; e domínio de tipos (datatype domain), que consiste em valores que pertencem a tipos de dados oriundos de XML Schema.

Horrocks et al. (2001, p.2) sugere:

> a separação entre tipos de dados e classes implica em que os primeiros acabam por ser modelados fora da ontologia, o que facilita não só a manutenção da simplicidade e controle de tamanho da linguagem de representação da ontologia, mas também facilita a implementação de seu suporte ao raciocínio.

A linguagem DAML+OIL provê meios para modelar domínios de conhecimento através de ontologias. DAML+OIL incorpora aspectos tanto da linguagem DAML quanto da linguagem OIL, vista por alguns como um subdialeto desta. Existem várias diferenças entre as linguagens OIL e DAML+OIL. A principal diferença é que a linguagem DAML+OIL foi baseada em RDF. Assim, é possível ver construções em RDF identificadas como DAML+OIL, mas não em OIL.

Até novembro de 2009 haviam 282 ontologias submetidas à biblioteca DAML, que podem ser verificadas em (http://www.daml.org/ontologies) e ainda uma lista de 243 ferramentas (validadores, navegadores, editores...etc.) relacionadas com a linguagem. A lista completa pode ser verificada em (http://www.daml.org/tools). A quantidade de ontologias e ferramentas desenvolvidas com DAML as credencia como linguagens das mais importantes para a construção de ontologias.

4.4.5 Web Ontology Language (OWL)

A OWL é uma linguagem de marcação semântica para a definição, instanciação, publicação e partilha de ontologias na World Wide Web. OWL é desenvolvida como uma extensão do vocabulário RDF (Resource Description Framework) e é proveniente de uma revisão das linguagens DAML + OIL. (BECHHOFER, 2004)

A linguagem OWL é reconhecida, atualmente, como o último padrão em linguagens para ontologia e recomendada como a principal linguagem para construção de ontologias, pelo consórcio W3C.

Apesar do alto investimento na criação das linguagens DAML e OIL e, posteriormente, DAML+OIL, o resultado ainda precisava de alterações, e a linguagem OWL foi originada justamente após se acrescentarem requisitos

de internacionalização e de documentação, como rótulos para axiomas, nomes locais únicos, entre outros.

A linguagem OWL tem como objetivo principal atender às necessidades de aplicação da Web Semântica e foi projetada para: construir ontologias, explicitar fatos sobre um domínio, definir indivíduos que fazem parte de um domínio e afirmações sobre ele, definir classe e propriedades destas classes, especificar como derivar consequências lógicas (fatos não literalmente presentes na ontologia, mas resultantes de sua semântica) e racionalizar sobre ontologias e fatos.

A OWL foi projetada com o objetivo de ser efetivamente utilizada por aplicações que necessitem processar o conteúdo de informações, e não somente apresentar a visualização destas informações.

Apesar de ser baseada em RDF e RDF Schema e utilizar-se da sintaxe XML, a linguagem OWL é considerada mais adaptada e mais fácil para expressar significados e semânticas que o conjunto XML, RDF e RDF Schema.

A linguagem OWL oferece três sublinguagens, projetadas para uso de implementadores e comunidades específicas, que se apresentam a seguir em ordem de expressividade: OWL Lite, OWL DL e OWL Full.

O OWL Lite dá suporte à criação de hierarquias simplificadas, que implementam restrições simples. Por ser mais simples e, consequentemente, apresentar uma gama menor de funcionalidades, é o mais utilizado na criação de ferramentas, portanto mais ferramentas suportam essa sublinguagem. A facilidade apresentada em relação ao OWL DL e ao OWL Full é uma de suas principais características, e o principal objetivo é fornecer um rápido caminho de migração para tesauros e outras taxonomias.

O OWL DL possui o mesmo vocabulário da linguagem OWL Full e dá suporte aos usuários que desejam o máximo de expressividade, sem perder a completude computacional (todas as conclusões são garantidas de serem computadas) e capacidade de decisão (todas as computações serão finalizadas em um tempo finito) dos sistemas de raciocínio. O OWL DL inclui

todos os construtores da linguagem OWL, com restrições, como separação entre tipos (uma classe não pode ser ao mesmo tempo um indivíduo ou tipo, e uma propriedade não pode ser ao mesmo tempo um indivíduo ou uma classe). OWL DL tem expressividade menor que o OWL Full, mas conta com melhor eficiência, computacionalmente falando, pois garante que todas as conclusões sejam computáveis (implementadas em máquinas que contenham processador) e que todas as computações sejam resolvidas num tempo finito. OWL DL tem esse nome devido a sua correspondência à Lógica de Descrição, ou Description Logic, um campo de pesquisa que tem estudado um fragmento de decisão particular de primeira ordem lógica.

O OWL Full foi desenvolvido para os usuários que desejam o máximo de expressividade e liberdade sintática do RDF, sem nenhuma garantia computacional. A linguagem OWL Full não conta com as restrições da OWL DL, e justamente por isso pode ser mais bem adaptada a situações onde o ponto mais importante é a expressividade. A OWL Full e a OWL DL suportam o mesmo conjunto de construções da linguagem OWL, embora com restrições um pouco diferentes. A OWL Full permite misturar OWL com RDF Schema e não requer a disjunção de classes, propriedades, indivíduos e valores de dados. Isto é, uma classe pode ser ao mesmo tempo uma classe e um indivíduo.

Segundo Harmelen e McGuinness (2004, p.4),

> a escolha de qual sub-linguagem OWL os desenvolvedores de ontologias devem usar vai depender das necessidades da ontologia. A escolha entre OWL Lite e OWL DL dependerá da necessidade das propriedades computacionais de OWL Lite ou das construções mais expressivas providas pela OWL DL. A escolha entre OWL DL e OWL Full dependerá da necessidade de expressividade, decidibilidade e completude computacional da OWL DL ou da expressividade e das facilidades da meta-modelo RDF Schema sem a previsibilidade computacional de OWL Full.

Portanto, o uso de uma ou outra especificação da linguagem OWL está diretamente ligada à análise prévia do domínio e do tipo de ontologia que será necessário criar.

4.4.5.1 Estrutura OWL – Namespaces

O início de um arquivo OWL tem como característica a declaração de namespaces no seu início. Os namespaces são responsáveis por fazer com que os indicadores que serão utilizados na ontologia sejam interpretados sem ambiguidade, pois através desta declaração é possível apenas sinalizar durante o conteúdo do arquivo o uso de vocabulários já pré-definidos.

A indicação do vocabulário empregado em cada termo garante que os termos utilizados na ontologia possam ser interpretados sem ambiguidade.

Conforme indica o W3C (2009), normalmente uma ontologia começa com uma declaração (exemplo 6): http://www.w3.org/TR/2004/REC-owl-guide-20040210/ #StructureOfOntologies

A segunda e a terceira linha representam a declaração dos namespaces desta própria ontologia.

```
<rdf:RDF
  xmlns     ="http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#"
  xmlns:vin ="http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#"
  xml:base  ="http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#"
  xmlns:food="http://www.w3.org/TR/2004/REC-owl-guide-20040210/food#"
  xmlns:owl ="http://www.w3.org/2002/07/owl#"
  xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd ="http://www.w3.org/2001/XMLSchema#">
```

EXEMPLO 6 – ESTRUTURA DE ONTOLOGIAS

A primeira declaração, que não tem um prefixo, indica que qualquer nome utilizado sem prefixo durante o desenvolvimento da ontologia será referenciado como da própria ontologia.

A segunda declaração indica a utilização do prefixo vin para referenciar uma ontologia de vinhos pré-definida. Esta ontologia sobre vinhos é exaustivamente citada dentro do contexto da Web Semântica e

muito referenciada dentro da definição da linguagem OWL no domínio da W3C.

A terceira declaração indica de onde foi constituída a base da nova ontologia e aponta que se utilizará uma ontologia já constituída (novamente a de vinhos) para iniciar a construção da nova ontologia. Esta declaração indica o reuso de uma informação já existente e demonstra a capacidade da linguagem de utilizar estruturas já prontas para constituir novas.

As próximas declarações indicam que, em alguns momentos, durante o desenvolvimento, serão utilizados os prefixos food, que também representa uma ontologia pré-definida, além dos prefixos owl, rdf, rdfs e xsd, que servem para indicar a utilização dos vocabulários referenciados, sinalizando a maneira que a ontologia será interpretada, ou seja, indicando o uso das primitivas já definidas e que são base para utilização da linguagem OWL.

Ressalta-se ainda que a utilização dos prefixos tenha como principal objetivo evitar a utilização da declaração completa das definições apresentadas, dando mais clareza ao código que está sendo desenvolvido.

4.4.5.2 Estrutura OWL – Cabeçalhos

Em seguida à definição dos namespaces, a definição de um arquivo OWL sugere um cabeçalho que indique um conjunto de informações a respeito da ontologia que está sendo desenvolvida.

É neste momento que deverão e poderão ser apontadas as informações que dão suporte a tarefas cruciais do desenvolvimento da ontologia, como: comentários, sinalização do controle de versão da ontologia, importação de um código já pré-existente, além da caracterização dos metadados referentes à ontologia a ser desenvolvida.

Estas informações devem ser agrupadas dentro da tag owl:Ontology, como verificado no código do exemplo 7.

```
<owl:Ontology rdf:about="">
  <rdfs:comment>Exemplo de Ontologia - CI</rdfs:comment>
  <owl:versionInfo> 07/07/2009 22:15:15 </owl:versionInfo>
  <owl:priorVersion rdf:resource="http://www.w3.org/TR/2003/PR-owl-guide-
20031215/wine"/>
  <owl:imports rdf:resource="http://www.w3.org/TR/2004/REC-owl-guide-20040210/food"/>
  <rdfs:label>Vinhos - Ontologia</rdfs:label>
  ...
```

EXEMPLO 7 – TAG OWL:ONTOLOGY

A tag inicial owl:Ontology, indica o local onde deverão ser apresentados os metadados para o documento a ser desenvolvido. Neste e em outros casos, a indicação desta tag não garante que será desenvolvida uma ontologia completa, podendo ser em alguns casos apenas a definição de algumas classes ou propriedades de um domínio, o que indicaria um arquivo complementar a uma ontologia.

O atributo rdfs:comment permite a indicação de comentários para a ontologia em desenvolvimento. Já os atributos owl:priorVersion e owl:versionInfo indicam, respectivamente, a última versão antes da que está em desenvolvimento, facilitando o processo de controle de versão e a versão da ontologia que está sendo desenvolvida. O atributo owl:imports permite a inserção de dados de outros arquivos dentro do documento que está sendo desenvolvido.

É importante ressaltar a diferença entre indicar o namespace para uma ontologia e a importação da mesma. A indicação de namespace ocorre quando se deseja utilizar parte da estrutura de outro documento, como definição de classes ou atributos, por exemplo. Já a utilização da tag owl:imports indica que o conteúdo completo da outra ontologia será inserido no seu documento.

A tag rdfs:label tem a função de nomear a ontologia que está sendo desenvolvida.

4.4.5.3 Elementos Básicos OWL – Classes

Entre os elementos fundamentais da linguagem OWL, destacam-se as classes, que são responsáveis por representar um grupo de indivíduos com características comuns, provendo um mecanismo de abstração para agrupar recursos com características similares, ou seja, as classes têm a característica de representar um conjunto ou uma coleção de indivíduos que compartilham das mesmas características.

A classe é utilizada para definir o conceito abstrato de um determinado domínio como pessoas, bichos, coisas, automóveis. São as raízes de uma taxonomia.

Segundo Bechhofer (2004, p.9),

> a linguagem OWL define como classe principal a classe owl:Thing, sendo assim, cada indivíduo na OWL é membro da classe owl:Thing. Deste modo, ela é superclasse de todas as classes OWL definidas pelos usuários. A linguagem OWL também apresenta a classe owl:Nothing, que indica que uma classe não possui instâncias, que é uma subclasse de todas as classes OWL. Uma classe é sintaticamente representada como uma instância nomeada da owl:Class, que é uma subclasse da rdfs:Class.

As classes em OWL podem ser definidas da seguinte maneira:

```
<owl:Class rdf:ID="Computador" />
<owl:Class rdf:ID="Fornecedor" />
<owl:Class rdf:ID="Esporte" />
```

EXEMPLO 8 – CLASSES OWL

O código de definição das classes Computador, Fornecedor e Esporte, apresentado, apenas indica a sintaxe de definição de uma classe, descritas através da tag owl:Class, com a indicação do atributo rdf:ID. Note-se que no exemplo 8 está apenas a definição da classe, que não tem validade nenhuma como ontologia. Para se completar uma ontologia, deve-se implementar as características que fazem parte dessa classe, como os indivíduos, as propriedades, a relação com outras classes. Mais adiante será

visto como ampliar o relacionamento entre classes, assim como inserir na ontologia informações sobre indivíduos e propriedades.

A construção de uma taxonomia só é possível através da definição de uma hierarquia de classes, que pode ser criada através da tag rdfs:subClassOf.

O exemplo 9 define uma hierarquia de classes:

```
<owl:Class rdf:ID="Notebook">
        <rdfs:subClassOf rdf:resource="#Computador"/>
...
</owl:Class>
```

EXEMPLO 9 – HIERARQUIA DE CLASSES

Esta declaração mostra que a classe Notebook é definida como uma subclasse da classe Computador, então, o conjunto de indivíduos da classe Notebook deve ser um subconjunto do conjunto de indivíduos da classe Computador.

Este tipo de construção permite construir uma frase como: "Notebook é um tipo de Computador", estabelecendo uma ligação "tipo-de".

A construção de uma classe também pode ser documentada através da tag rdfs:comment. Outras definições também podem ser utilizadas na criação de uma classe, como a tag owl:disjointWith, que indica que uma classe não pode compartilhar instâncias com classe que tem esse tipo de relacionamento, conforme será visto no exemplo 10, que faz referência a pratos com carne e vegetarianos.

```
<owl:Class rdf:ID="Vegetarianos">
</owl:Class>
```

EXEMPLO 10 – CONSTRUÇÃO DE UMA CLASSE

4.4.5.4 Elementos Básicos OWL – Indivíduos

Indivíduos são definidos como objetos do mundo que sempre estão ligados às classes previamente definidas, ou seja, são membros das classes. Os indivíduos podem e devem estar ligados a outros indivíduos e são caracterizados através dos valores atribuídos as suas propriedades.

Para inserir um indivíduo em uma ontologia definida em OWL, é necessário apresentá-lo como membro de uma classe (exemplo 11).

```
<Notebook rdf:ID="Tablet" />
```

EXEMPLO 11 – INDIVÍDUO

A definição apresentada no exemplo 11 indica uma das possibilidades de se declarar um indivíduo chamado Tablet. Esta construção indica que o indivíduo Tablet é uma instância da classe Notebook, declarando um fato sobre a ontologia Computador, implicando em afirmar que "Tablet é um Notebook". Além da declaração apresentada no exemplo 11, pode-se também definir um indivíduo com o conjunto de linhas apresentadas no exemplo 12.

```
<owl:Thing rdf:ID="Tablet" />

<owl:Thing rdf:about="#Tablet">
  <rdf:type rdf:resource="#Notebook" />
<owl:Thing>
```

EXEMPLO 12 – OUTRO EXEMPLO DE INDIVÍDUO

4.4.5.5 Elementos Básicos OWL – Propriedades

Propriedades são recursos da linguagem OWL que têm o propósito de descrever fatos em geral. As propriedades são utilizadas para estabelecer relacionamentos entre os indivíduos ou ainda entre indivíduos e valores. Através das propriedades, pode-se fazer referência a todos os membros de uma classe, ou seja, afirmar fatos gerais sobre os membros de uma classe ou

então a apenas um indivíduo específico de determinada classe. As propriedades em OWL são relacionamentos binários.

A linguagem OWL define duas categorias principais para propriedades:

- Propriedades de objetos (object properties): estabelece relação entre indivíduos ou classes.
- Propriedades de dados (datatype properties): que indicam a relação entre indivíduos, que são instâncias das classes, e valores de dados expressos em RDF e tipos do XML Schema. O W3C recomenda, através do endereço (http://www.w3.org/TR/2004/REC-owl-guide-20040210 /#SimpleProperties) no item 3.3.2, um conjunto de tipos definidos em XML Schema para utilização da linguagem OWL.

Qualquer propriedade definida em um documento OWL é subclasse da classe RDF rdf:Property. Propriedade de objetos é definida como instância de classe owl:ObjectProperty, e propriedade de dados é definida como instância da classe owl:DatatypeProperty.

```
<owl:ObjectProperty rdf:ID="endereco">
   <rdfs:label>Endereço</rdfs:label>
   <rdfs:domain rdf:resource="#Fornecedor"/>
   <rdfs:range rdf:resource="#CEP"/>
 </owl:ObjectProperty>
```

EXEMPLO 13 – PROPRIEDADE DE OBJETOS

O exemplo 13 apresenta a definição de uma propriedade de objetos, indicando que a classe Fornecedor, tem uma propriedade denominada Endereço, que deve ser obrigatoriamente preenchida com valores da classe CEP. A classe CEP já deve existir na ontologia.

```
<owl:DataTypeProperty rdf:ID="qtdeProcessadores">
  <rdfs:domain rdf:resource="Computador" />
  <rdfs:range rdf:resource="&xsd;positiveInteger"/>
</owl:DatatypeProperty>
```

EXEMPLO 14 – PROPRIEDADE DE DADOS

O exemplo 14 apresenta a definição de uma propriedade de dados, indicando que a classe Computador, já definida anteriormente, tem a propriedade qtdeProcessadores, e que esta propriedade só aceita inteiros positivos, de acordo com a definição &xsd;positiveInteger, que é um tipo de dado previamente definido.

```
<owl:DatatypeProperty rdf:ID="rua">
   <rdfs:label>Rua, Avenida ou Logradouro</rdfs:label>
   <rdfs:subPropertoyOf rdf:resource="#endereco"/>
   <rdfs:domain rdf:resource="#Fornecedor"/>
   <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
```

EXEMPLO 15 – SUB-PROPRIEDADE OWL

Assim como as propriedades, uma ontologia OWL pode definir subpropriedades que são propriedades da propriedades. O exemplo 15 refere-se a uma subpropriedade Rua que está definida como uma subpropriedade da propriedade Endereço. Neste caso, a subpropriedade Rua é definida como uma informação do tipo string e que faz parte do domínio da classe Fornecedor.

```
<Notebook rdf:ID="Tablet">
  <giroTelaGraus rdf:datatype="xsd;positiveInteger">180</giroTelaGraus>
</Notebook>
```

EXEMPLO 16 – SUB-PROPRIEDADE DE DADOS APLICADA A INDIVÍDUO

As propriedades de dados também podem ser aplicadas diretamente a indivíduos quando forem específicas de uma instância apenas da classe, conforme é demonstrado no exemplo 16, que define em 180 graus o giro da tela de um Tablet, que é um indivíduo da classe Notebook.

Uma boa definição das propriedades e subpropriedades de uma classe é fundamental para que se tenha uma boa qualidade na definição da ontologia.

4.4.5.6 Elementos Básicos OWL –Restrições em Propriedades

A linguagem OWL utiliza as propriedades para impor restrições na definição de uma ontologia. Uma restrição é uma imposição de limites que uma determinada classe ou indivíduo deve seguir. As restrições apresentadas pela linguagem OWL podem ser de dois tipos:

- Restrições de Cardinalidade
- Restrições de Valores.

A utilização de restrições de cardinalidade está diretamente ligada a permitir que uma instância de uma classe possa ter um número arbitrário de valores para uma determinada propriedade. Segundo Bechhofer et al.(2009, p.12), a OWL provê três construções para cardinalidade:

- owl:maxCardinality: descreve uma classe de todos os indivíduos que têm, no máximo, N valores semanticamente distintos.
- owl:minCardinality: descreve uma classe de todos os indivíduos que têm, no mínimo, N valores semanticamente distintos. Esta restrição é um meio para dizer que uma propriedade requer um valor para todas as instâncias da classe.
- owl:cardinality: descreve uma classe de todos os indivíduos que têm exatamente N valores semanticamente distintos.

```
<owl:DataTypeProperty rdf:ID="qtdeProcessadores">
  <rdfs:domain rdf:resource="Computador" />
  <owl:Restriction>
        <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
  </owl:Restriction>
</owl:DatatypeProperty>
```

EXEMPLO 17 – RESTRIÇÃO DE CARDINALIDADE

O exemplo 17 apresenta uma restrição de cardinalidade mínima, que é referenciada no código através das tags owl:Restriction e owl:minCardinality, dando à propriedade qtdeProcessadores a necessidade de apresentar minimamente o valor 1. Dependendo da necessidade do

código, poderiam ser utilizadas as tags owl:cardinality ou owl:maxCardinality, ao invés de owl:minCardinality.

As restrições de valores se dividem em três tipos: allValuesFrom, someValuesFrom e hasValue, e têm como principal característica o fato de serem restrições locais, diferentes das restrições domain e range, que são globais.

Os recursos apresentados nesta pesquisa sobre linguagem OWL permitem iniciar o desenvolvimento de ontologias e entender um pouco sobre o conceito de desenvolvimento, visto que a linguagem apresenta recursos de várias outras linguagens e assemelha-se à metodologia de desenvolvimento Orientado a Objetos, utilizada em linguagens de programação. O guia da linguagem, disponível no site do W3C, apresenta muitos outros recursos não citados aqui, porém contribuirão para aumentar a complexidade e as funcionalidades de uma ontologia escrita em OWL.

4.5 Ferramentas para desenvolvimento de ontologias

Com base em estudos de FARQUHAR, FIKES, RICE (1997), apresentam-se várias metodologias para a construção de ontologias.

As metodologias apresentadas possuem abordagens e características diversas. Para verificar a utilidade das metodologias e utilizar uma base de comparação, é necessário avaliar os resultados da aplicação de cada uma.

Além de metodologias, existem ferramentas utilizadas para a construção de uma ontologia.

O desenvolvimento de uma ontologia pode ser realizado através de um editor de texto puro, escrevendo-se o código como se estivesse desenvolvendo um programa, porém o uso de ferramentas para auxiliar no processo de construção de ontologias é muito bem-vindo, visto que ele contribui na agilidade de desenvolvimento e minimiza os erros,

principalmente os de sintaxe. A seguir serão apresentadas algumas ferramentas utilizadas para o desenvolvimento de ontologia, priorizando e dando maior ênfase às ferramentas que oferecem recurso para desenvolvimento com a linguagem OWL, dada a indicação da W3C em relação à linguagem.

4.5.1 OilEd[18]

Um dos editores mais simples encontrados é o OilEd. Teve como objetivo inicial estimular o interesse pela linguagem DAML + OIL. Apresenta limitações para o desenvolvimento de ontologias em larga escala, não suportando versionamento, migração e integração de ontologias existentes, argumentação e outras tarefas do processo de construção de ontologias.

O OilEd suporta linguagem OWL e é freeware. O plug-in OilViz pode ser incorporado ao software, permitindo uma visualização mais rica da estrutura de classes da ontologia que o visualizador que vem inicialmente incorporado ao OilEd.

O projeto OilEd prevê que o software seja disponibilizado em uma base open source e adquira licença GPL em um futuro próximo, com o objetivo de ter seu código melhorado pela comunidade científica.

4.5.2 OntoEdit[19]

OntoEdit é um ambiente gráfico de desenvolvimento e edição de ontologias que segue os padrões do W3C e permite inspeção, codificação, navegação e alteração de ontologias, inclusive com suporte a exportação de ontologias em tecnologias como: RDF(S), XML e DAML+OIL.

---

[18] http://img.cs.man.ac.uk/oil/
[19] http://www.ontoprise.de

A versão disponibilizada, shareware, possibilita o desenvolvimento de ontologias com um número limitado de conceitos. Para usufruir de todos os recursos da ferramenta, é necessária a aquisição da licença comercial.

O editor que faz parte do projeto On-To-Knowledge implementa um processo específico para a construção de ontologias em três fases: requisitos que descrevem as atividades, refinamento da ontologia de acordo com a aplicação e a fase de avaliação. Cada fase usa ferramentas integradas ao ambiente, de acordo com suas características específicas.

O OntoEdit não permite desenvolvimento de ontologias em linguagem OWL.

4.5.3 Chimaera[20]

O Chimaera tem uma característica diferente das ferramentas apresentadas até este momento, pois seu objetivo principal é resolver diferenças entre ontologias diferentes, portanto tem a função de ser uma ferramenta de diagnóstico de ontologias, para verificação de sintaxe, comparação de ontologias, indicando classes e atributos semelhantes.

A ferramenta Chimaera pode ser utilizada como auxiliar no desenvolvimento de ontologias, principalmente porque pode combinar ontologias unindo classes ou ainda criando uma hierarquia de classe e subclasse entre classes semelhantes de ontologias diferentes, além de resolver conflitos de nomes e reorganizar de forma taxonômica a ontologia.

A ferramenta está disponível online no site da Universidade de Stanford através do link (http://www.ksl.stanford.edu/software/chimaera/), e permite login como usuário cadastrado ou anônimo.

Segundo o site, a ferramenta pode carregar e exportar resultados em DAML e OWL, além de uma gama enorme de outras linguagens para desenvolvimento de ontologias.

4.5.4 API Jena[21]

A API Jena é um framework desenvolvido com o objetivo de auxiliar no desenvolvimento de aplicativos para Web Semântica. O framework foi inicialmente desenvolvido nos laboratórios da HP e tem como principal característica um mecanismo de inferência associado ao suporte das linguagens RDF, RDF Schema e OWL.

A API permite o desenvolvimento e manipulação de ontologias através de softwares que utilizam linguagem orientada a objetos, como Java, por exemplo.

A ferramenta é freeware e está disponível para download.

4.5.5 Protégé 2000[22]

Protégé 2000 é um ambiente open source para: criação e edição de ontologias e bases de conhecimento.

A plataforma Protégé suporta dois tipos de modelagem para o desenvolvimento de ontologias: o Protégé-Frames e o Protégé-OWL. As ontologias desenvolvidas no Protégé podem ser exportadas para vários formatos, entre eles RDF, RDF Schema, OWL e XML.

O ambiente Protégé é baseado em Java, é extensível, e fornece uma estrutura que permite aos desenvolvedores de todo mundo a ampliação do software através do desenvolvimento de plug-ins.

O Protégé 2000 permite a construção de ontologias de domínio, combinação/integração de ontologias existentes e o armazenamento de uma base de conhecimento sobre determinado domínio.

A interface gráfica apresentada em sua versão desktop (figura 21) é bem intuitiva para usuários que já conhecem a estrutura de desenvolvimento de

---

[20] http://www-ksl.stanford.edu/software/chimaera/
[21] http://jena.sourceforge.net/ontology/
[22] http://protege.stanford.edu/

ontologias. A novidade atual refere-se à versão Alpha Web Protege, que permite a utilização da ferramenta diretamente de um browser Web.

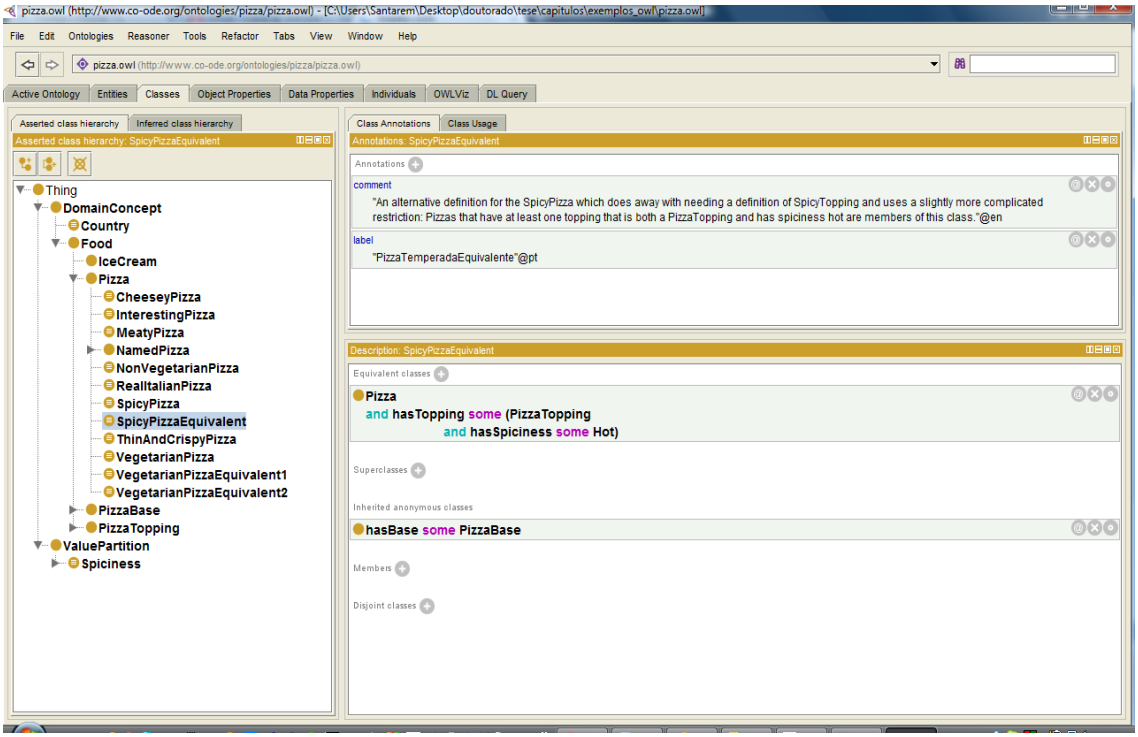


Figura 21 – Protégé 2000
Fonte: Próprio autor

Todas as relações apresentadas na linguagem OWL podem ser implementadas utilizando-se o Protégé 2000.

O Protégé 2000 foi desenvolvido, inicialmente, para atender às necessidades de ontologias médicas, através do Departamento de Informática Médica da Universidade de Stanford, tendo como projeto inicial uma ferramenta de aquisição de conhecimento para um sistema especialista para oncologia.

A ferramenta passou a adotar a filosofia de código aberto, a partir do momento em que foram verificadas as potencialidades de desenvolvimento que a arquitetura inicial do projeto disponibilizava. A partir do desenvolvimento do código, o Protégé efetivou sua evolução, principalmente na apresentação gráfica de ontologias.

O Protégé 2000 destaca-se entre as ferramentas open source disponíveis para desenvolvimento e manipulação de ontologias, especialmente pela apresentação visual clara e pela facilidade de operação para usuários inexperientes.

4.5.6 Outras iniciativas

Além das ferramentas descritas, há outras iniciativas de desenvolvimento de ferramentas ao redor do mundo, dentre as quais se destacam: Ontokem, Ontoeditor, CODEA, WebODE, OntoEdit, KAON, JOE.

4.6 Construção Automática de Ontologias

Se por um lado a construção de ontologias é vista com bons olhos no objetivo de auxiliar a construção de uma rede semântica de informações, existe outro lado, que são os conjuntos de informações já armazenados e que poderiam contribuir para o desenvolvimento de uma ontologia baseada no volume de dados cadastrados.

Como já foi visto, a construção de ontologias despende um processo bastante longo e complexo de aquisição do conhecimento sobre o domínio a ser desenvolvido e, dessa forma, construir uma ontologia sobre um conjunto de informações previamente cadastradas, que, na maioria das vezes, se apresenta de forma pouco estruturada, podendo demandar um trabalho de muito tempo.

Alguns casos são reconhecidamente conjuntos de informações bastante ricos, como: prontuários médicos, sistemas de gestão da informação como os ERP's, boletins de ocorrências policiais, dados semiestruturados, dicionários, entre outros, e que podem conduzir para a sistematização de uma ontologia. Porém fica claro que a recuperação destas

informações, para que a ontologia seja realizada por seres humanos, tem um nível de complexidade bastante alto, visto o nível de subjetividade empregado neste processo.

Alguns estudos têm conduzido para a utilização de técnicas e métodos que possam minimizar o tempo para construir, melhorar ou ainda atualizar ontologias de forma automática, utilizando-se bases de conhecimento já estabelecidas, como as já referenciadas.

Várias pesquisa tendem a aperfeiçoar os métodos de desenvolvimento automático de ontologias, porém alguns itens são constantemente citados: a fidelidade da fonte a partir da qual se está construindo a ontologia e também as relações implícitas que existem em textos, como livros, jornais e artigos.

Notadamente, o trabalho de desenvolvimento de ontologias através de técnicas automáticas não tem apresentado resultados efetivamente seguros, porém tem contribuído para o desenvolvimento de ontologias, de forma que, com a interferência humana em praticamente todas as fases do processo de geração da ontologia, possa construir uma estrutura inicial de classes e indivíduos e, posteriormente, ser analisado e modificado novamente por interferência humana.

Segundo Mayrink e Ladeira (2008, p.5),

> [...] é de extrema importância a presença de um especialista durante algumas fases do desenvolvimento, principalmente durante a aquisição de conhecimento e validação da heurística criada, sendo que esse pode sugerir categorias a serem implementadas e verificar se as mesmas estão apropriadas após a extração de termos. No caso da heurística utilizada, ele poderia recomendar quais as expressões seriam empregadas na identificação dos termos a serem extraídos.

Nas abordagens sobre geração automática de ontologias, a partir de uma base de conhecimento, fica claro que existe muito trabalho a ser desenvolvido com objetivo de alcançar resultados que possam efetivamente ser utilizados sem interferência humana, porém a criação de técnicas e métodos tem contribuído no sentido de colaborar na construção de

ontologias e minimizado o volume de trabalho que seria inicialmente realizado.

4.7 Ontologias de Topo

Atualmente, existem diversos esforços no sentido de construção de ontologias de topo, isto é, aquelas cujo objetivo é representar o conhecimento humano e servir como referência básica para construção de ontologias de domínio e de aplicação.

Entre os projetos mais conhecidos, destacam-se as ontologias Sumo, KR e projeto CYC.

A Ontologia SUMO (Suggested Upper Merged Ontology) foi proposta por um grupo de trabalho da IEEE, formada por colaboradores de diversas áreas, com a finalidade de oferecer uma ontologia com termos genéricos entre 1000 e 2500 termos.

A Ontologia KR (Knowledge Representation), proposta por John Sowa, é baseada em abordagens filosóficas e, principalmente, na semiótica de Peirce e categorias de existência enumeradas por Whitehead.

O projeto CYC, que representa o maior esforço no sentido de desenvolver uma ontologia com a maior amplitude possível, é projetado para atender todo o conhecimento humano e apresenta cerca de 3000 classes superiores, divididas em 43 categorias, contando com aproximadamente 2.000.000 de conceitos.

A ontologia CYC teve sua origem em 1984, através Doug Lenat, da Microelectronics and Computer Corporation, MCC, sendo proprietária e restrita. Hoje os direitos são detidos pela Cycorp.

A Cycorp apresenta uma versão gratuita disponível para uso e consulta chamada OpenCyc, considerado como um padrão pelo IEEE.

```
<rdf:RDF xml:base="http://sw.opencyc.org/concept/"
    xmlns="http://sw.opencyc.org/concept/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
```

```
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:cyc="http://sw.cyc.com/"
xmlns:opencyc="http://sw.opencyc.org/"
xmlns:cycAnnot="http://sw.cyc.com/CycAnnotations_v1#">

<owl:Ontology rdf:about="">
    <owl:versionInfo>Version 2.0.0</owl:versionInfo>
    <rdfs:comment xml:lang="en">
 OpenCyc Knowledge Base

 Copyright© 2001-2009 Cycorp, Inc., http://www.cyc.com/, Austin, TX, USA

 This file contains an OWL representation of information contained
 in the OpenCyc Knowledge Base. The content of this OWL file is
 licensed under the Creative Commons Attribution 3.0 license whose
 text can be found at http://creativecommons.org/licenses/by/3.0/legalcode.
 The content of this OWL file, including the OpenCyc content it represents,
 constitutes the "Work" referred to in the Creative Commons license. The terms of
 this license equally apply to, without limitation, renamings and other
 logically equivalent reformulations of the content of this OWL file
 (or portions thereof) in any natural or formal language, as well
 as to derivations of this content or inclusion of it in other ontologies.

 Mappings between OpenCyc terms and Wikipedia article names provided by
 Olena Medelyan and Catherine Legg, University of Waikato, NZ under a Creative
 Commons Attribution 3.0 license.

    </rdfs:comment>
</owl:Ontology>
```

EXEMPLO 18 - CABEÇALHO EM OWL DA ONTOLOGIA OPENCYC

A versão OpenCyc, que atualmente se apresenta no release 1.0.2, tem as seguintes características:

- Versão gratuita com 47.000 conceitos
- 306.000 sentenças sobre conceitos
- Parte do Mecanismo de inferência do Cyc
- Browser para visualização da Ontologia
- OWL e Cyc (Linguagem própria similar ao LISP)
- SubL (Interpretador: navegar/editar/inferir)

- API's para desenvolvimento de aplicações

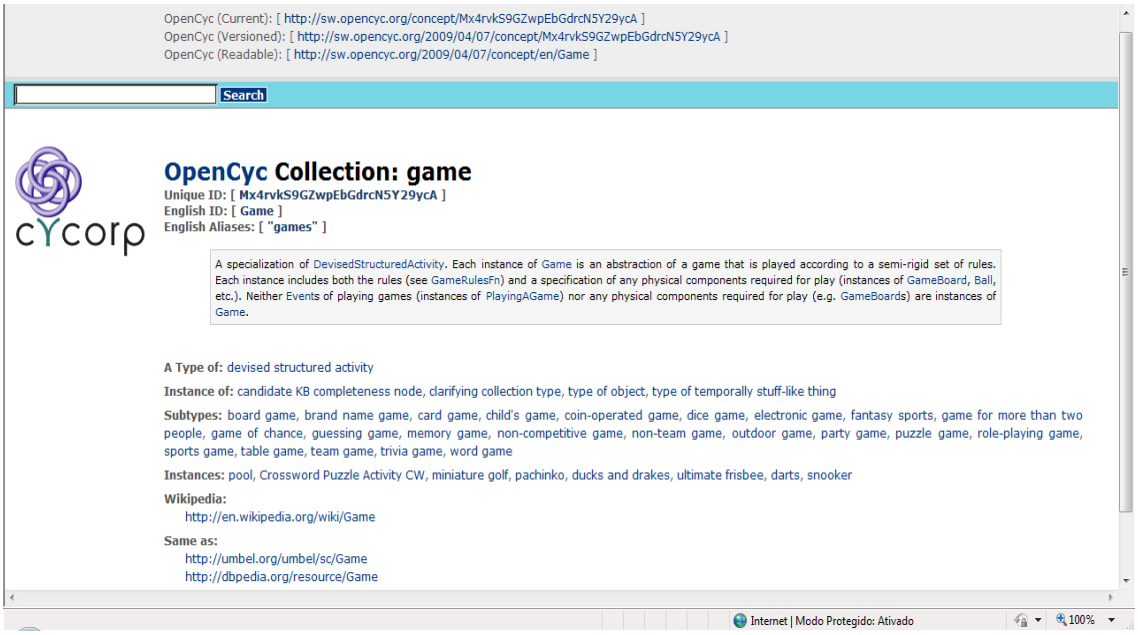


Figura 22 – OpenCyc
Fonte: http://sw.opencyc.org

Através do site do projeto, é possível baixar a versão da OpenCyc em formato OWL (exemplo 18), e também utilizar a ferramenta disponível para uso via web (figura 22).

As ontologias de topo podem ser utilizadas diretamente ou então servir como base para construção de ontologias de domínio.

Este capítulo apresentou os conceitos, linguagens e ferramentas para construção e manipulação de ontologias, que são fundamentais no desenvolvimento e aplicação das técnicas da Web Semântica, finalizando a contextualização e base teórica a respeito da Web 3.0.

5 REPOSITÓRIOS DIGITAIS DE INFORMAÇÃO CIENTÍFICA

Nesse capítulo busca-se apresentar os repositórios digitais de informação científica, objeto principal desta pesquisa. Tem por objetivo conceituar os repositórios digitais e sua estrutura de informação, observando as principais ferramentas para a implementação de repositórios digitais disponíveis em formato open source. Através de observação direta não participativa também é feita uma análise e tabulação dos recursos de Web 2.0 e Web 3.0, implementados nas ferramentas para construção de repositórios digitais.

A produção de material científico e, principalmente, de revistas científicas cresceu significativamente durante todo o século XIX, em função do aumento do número de pesquisadores e de pesquisa. Durante o século passado, o crescimento foi significativo, incrementado pelo fato de as revistas científicas serem também publicadas por universidades e pelo Estado, além das editoras comerciais.

O aumento da quantidade de pesquisadores e, consequentemente, de pesquisas resultaram no aumento da demanda em relação ao acesso ao material científico já produzido, para que o processo de geração de conhecimento através de conhecimento já produzido fosse possível. Com o mercado de publicação científica em plena expansão, as comunidades científicas, que produzem material, passaram a ter dificuldade de acesso à produção, visto que o conteúdo passou a ser gerido e explorado pelas editoras.

Se, por um lado, a maior parte dos periódicos científicos relevantes internacionalmente é distribuída por agentes comerciais que, por terem o direito de venda exclusiva da publicação, adotam preços elevados, por outro, observa-se que os produtores científicos mantêm uma competição com os editores comerciais, produzindo revistas constituídas com o objetivo de divulgar a sua própria produção científica, sem terem que abrir mão de seus direitos autorais para os editores (LEVACOV, 1997).

A dificuldade e a necessidade de acesso ao material já produzido, juntamente com a introdução da tecnologia digital, estabeleceram uma nova ordem na edição e publicação da comunicação científica: o surgimento das publicações científicas em meio eletrônico e a aproximação e interação da comunidade científica, pela web, em novas escalas de tempo e espaço, além da criação de um novo conceito de publicação — o Open Access Initiative (OAI) — que tem como premissa promover o acesso livre e irrestrito à literatura científica e acadêmica, de forma a mudar a maneira de explorar o material científico produzido.

O OAI estabeleceu novos critérios em relação à maneira com que as instituições e os pesquisadores lidam com o material produzido em seu âmbito, porém o estabelecimento desta filosofia está amparado por estruturas tecnológicas que permitem a publicação e consequente disseminação da informação. Essas estruturas tecnológicas são encabeçadas, principalmente, pelas ferramentas que permitem a criação de Repositórios Digitais Institucionais e Revistas Eletrônicas.

Moreno, Leite e Arellano (2006, p.84) afirmam:

> Os arquivos/repositórios de acesso livre, baseados em arquivos abertos, são interoperáveis e, por esta razão, podem ser acessados por diversos provedores de serviços disponíveis em nível nacional e internacional. Dessa forma, os periódicos eletrônicos, os repositórios institucionais e os repositórios temáticos de acesso livre, aliados à tecnologia de arquivos abertos estão sendo utilizados pelas comunidades científicas para apoiar e tornar mais ampla a divulgação dos resultados das pesquisas bem como maximizar o seu impacto, criando mecanismos para legitimar e estimular a publicação dos trabalhos produzidos.

Repositórios são conjuntos de documentos coletados, organizados e disponibilizados eletronicamente. No contexto específico dos repositórios, os documentos adquirem novas configurações e são denominados objetos digitais ou estrutura de dados digitalmente codificados, composta pelo conteúdo de informação, metadados e identificador (BEKAERT; VAN DE SOMPEL, 2006).

Os repositórios institucionais inserem-se no movimento conhecido por Open Access Initiative, que visa promover o acesso livre e irrestrito à literatura científica e acadêmica, favorecendo o aumento do impacto do trabalho desenvolvido pelos investigadores e instituições, e contribuindo para a reforma do sistema de comunicação científica. (RODRIGUES, 2006)

Nos últimos anos, os repositórios institucionais têm sido alvo de grande atenção por parte de universidades e bibliotecas universitárias, reassumindo o controle acadêmico sobre a publicação, aumentando a competição e reduzindo o monopólio das revistas científicas das editoras comerciais.

Os repositórios digitais são sistemas de informação que facilitam a publicação e o armazenamento de documentos, além de fornecer serviços de informação, e por isso o interesse em contribuir com a organização de sua informação.

As comunidades científicas, de um modo geral, têm visto a utilização de repositórios institucionais como um divisor de águas entre as formas de publicar trabalhos científicos, assim como disseminá-los entre pares e pesquisadores. Esses novos formatos são caracterizados, principalmente, pelo formato eletrônico de publicação, impulsionados pelas dificuldades encontradas na publicação impressa e pelo avanço tecnológico.

Os novos modelos de publicação científica, especialmente os ligados à publicação científica eletrônica, têm como premissa a quebra de algumas barreiras, como tempo, facilidade de publicação e disseminação dos trabalhos publicados.

Neste sentido, encontra-se, em universidade e institutos de pesquisa, um movimento em busca da facilidade de publicação e da utilização da Internet como meio de disseminar as pesquisas, sejam elas no ambiente acadêmico ou não.

As grandes universidades brasileiras, em especial as públicas, que contam com programas de pós-graduação, cumprindo solicitação da CAPES, já têm ou procuram iniciativas que buscam publicar pelo menos as

dissertações de mestrado e as teses de doutorado de maneira eletrônica, tornando de conhecimento público os trabalhos desenvolvidos.

Paralelo a esta frente de publicação dos trabalhos já defendidos, alguns órgãos de pesquisa têm também se esforçado no sentido de desenvolver ambientes de repositórios institucionais e temáticos, para publicação e autoarquivamento da pesquisa de sua comunidade.

Um repositório digital é uma forma de armazenamento de objetos digitais que tem a capacidade de manter e gerenciar material por longos períodos de tempo e prover o acesso apropriado. Essa estratégia foi possibilitada pela queda nos preços no armazenamento, pelo uso de padrões como o protocolo de coleta de metadados da Iniciativa dos Arquivos Abertos (OAI-PMH), e pelos avanços no desenvolvimento dos padrões de metadados que dão suporte ao modelo de comunicação dos arquivos abertos (VIANA, 2007).

Segundo Leite (2009, p. 21),

> Um repositório institucional de acesso aberto constitui, portanto, um serviço de informação científica – em ambiente digital e interoperável – dedicado ao gerenciamento da produção intelectual de uma instituição.
>
> Contempla, por conseguinte, a reunião, armazenamento, organização, preservação, recuperação e, sobretudo, a ampla disseminação da informação científica produzida na instituição. Uma das definições mais conhecidas é que um repositório institucional consiste em um conjunto de serviços que a universidade oferece para os membros da sua comunidade com vistas ao gerenciamento e disseminação do material digital criado pela instituição e pelos seus membros.

Os repositórios digitais podem ser divididos em temáticos e institucionais, além de apresentarem estrutura e características próprias.

O primeiro tipo de repositório digital, o repositório temático (RT), armazena documentos com uma delimitação de cobertura por assunto, área do conhecimento ou temática específica.

Kuramoto (2006, p. 83) define repositórios temáticos como “um conjunto de serviços oferecidos por uma sociedade, associação ou

organização, para gestão e disseminação da produção técnico-científica em meio digital, de uma área ou subárea específica do conhecimento". O êxito dos repositórios temáticos suscitou discussões sobre seu funcionamento e a necessidade de um gestor que lhes garantisse bom desempenho, fazendo surgir a figura de uma instituição responsável e agregadora das iniciativas individuais de desenvolvimento de repositórios. Neste momento emergem os repositórios institucionais (CAFÉ, 2003).

O repositório institucional (RI) é a reunião de repositórios temáticos, sob a responsabilidade técnica e administrativa de uma instituição ou organismo. Por consequência, este tipo de repositório é multidisciplinar e possui uma gama de tipos de documentos ainda maior que um repositório temático. Além de agregar o conjunto de informações relativas e/ou de interesse para a instituição, dispõem de serviços referentes à organização, disseminação e acesso ao conteúdo digital (CAFÉ, 2003).

Os repositórios digitais, sejam eles temáticos ou institucionais, apresentam características semelhantes, possuem uma estrutura comum de submissão e acesso às informações e são desenvolvidos segundo padrões de interoperabilidade específicos, que potencializam o uso desses sistemas para agregação e divulgação da informação digital. Algumas das características ou observações quanto à estrutura destes repositórios são impostas pelos padrões que utilizam.

De acordo com os objetivos propostos, um repositório digital pode contemplar uma infinidade de tipos de documentos, ou seja, ter uma tipologia variada de documentos. Mesmo com a concepção de que os repositórios foram desenvolvidos para divulgar documentos já publicados nos meios tradicionais (BUDAPEST..., 2002), não há uma delimitação sobre os tipos de documentos que atualmente podem compor um repositório (CAFÉ, 2003).

Os repositórios institucionais têm sido mais amplamente desenvolvidos em ambientes universitários com a preocupação de disponibilizar resultados de pesquisa a partir de coleções digitais de

departamentos e faculdades. Os repositórios dão projeção à produção acadêmica e a reúnem em um sistema de informação que possibilita seu acesso em longo prazo, com um custo inferior à publicação em revistas tradicionais, evidenciando seu valor científico, cultural, social e econômico (CROW, 2002).

Kuramoto (2006, p.101) afirma:

> Em muitos países, inclusive aqueles mais desenvolvidos, as agências de governo vêm elaborando e implantando ações em prol do acesso livre à informação. Pelo ROAR, verifica-se que países como os EUA,o Reino Unido e a Alemanha vêm investindo na construção de repositórios, despontando como os países que mais implantaram repositórios. Portanto, esses países servem de referência para as nossas ações concernentes a essa questão.

A implantação de um repositório institucional em uma determinada comunidade se inicia a partir de uma política de publicação de autoarquivamento, que indica a conscientização da necessidade de se criar uma cultura de postagem, passando pela implementação técnica do repositório e finalizando com a efetiva utilização do mesmo dentro da comunidade científica.

Leite (2009, p. 22) expressa a importância da utilização de repositórios digitais em ambientes acadêmicos.

> Instituições acadêmicas no mundo inteiro utilizam repositórios institucionais e o acesso aberto para gerenciar informação científica proveniente das atividades de pesquisa e ensino e oferecer suporte a elas. Nesse sentido, os repositórios institucionais têm sido intensamente utilizados para:
>
> - melhorar a comunicação científica interna e externa à instituição;
> - maximizar a acessibilidade, o uso, a visibilidade e o impacto da produção científica da instituição;
> - retroalimentar a atividade de pesquisa científica e apoiar os processos de ensino e aprendizagem;
> - apoiar as publicações científicas eletrônicas da instituição;

- contribuir para a preservação dos conteúdos digitais científicos ou acadêmicos produzidos pela instituição ou seus membros;
- contribuir para o aumento do prestígio da instituição e do pesquisador;
- oferecer insumo para a avaliação e monitoramento da produção científica;
- reunir, armazenar, organizar, recuperar e disseminar a produção científica da instituição.

Para a implementação técnica, são vários os softwares disponíveis, tanto em iniciativas de software livre, open-source e até alguns que sugerem a aquisição de uma licença de uso. Entre os principais, atualmente encontram-se: Dspace, GNU E-prints[23], OPUS[24], Open Repository[25], DiVA[26], Fedora[27].

Nesta pesquisa, todos os testes realizados e sugestões abordadas utilizam como parâmetro principal a ferramenta Dspace, por oferecer um ambiente altamente configurável, que pode ser empregado tanto para o desenvolvimento de pequenos repositórios até em ambientes complexos de tramitação de material científico avaliado por pares. O Dspace nasceu de um esforço conjunto de investigação do MIT[28] (Massachusetts Institute of Technology) e da HP (Hewlett-Packard), com sua primeira versão disponibilizada em novembro de 2002.

Cabe ressaltar que o Dspace está sendo abordado apenas como ferramenta auxiliadora no processo de construção de um modelo que tem como principal objetivo atender, em seu contexto tecnológico e conceitual, todos os tipos de ferramentas que permitam a constituição de repositórios digitais informacionais. Portanto, algumas características técnicas estarão direcionadas ao Dspace, porém poderão ser facilmente adaptadas, quando não compatíveis, com qualquer outra ferramenta.

---

[23] http://www.eprints.org/software/
[24] http://opus.bath.ac.uk/
[25] http://www.openrepository.com/
[26] http://www.diva-portal.org
[27] http://www.fedora-commons.org/
[28] http://web.mit.edu/

Utilizado como base para a implementação de repositórios institucionais, o Dspace facilita o processo de desenvolvimento dos mesmos, tanto na questão técnica quando na questão econômica. Por se tratar de um software, cujo modelo de licenciamento é o BSD Open Source License, não é necessário investimento financeiro na aquisição do software, incluindo ainda a possibilidade de as instituições de pesquisa criarem grupos que possam colaborar com o desenvolvimento da ferramenta. Outro fator importante da utilização do Dspace é a grande disseminação desta ferramenta ao redor do mundo, permitindo aos usuários e administradores de sistemas a troca de informações quanto à utilização e administração do sistema. Está atualmente em funcionamento no MIT, e em diversas universidades e outras instituições dos Estados Unidos e da Europa. O Dspace.org também propõe um ambiente que agrega vários colaboradores e desenvolvedores do mundo todo em prol de melhorias, tanto no desenvolvimento quanto no uso da ferramenta. É possível verificar os repositórios implementados com a ferramenta através do site o oficial do Dspace (http://www.dspace.org/content/view/1047/333/).

O Dspace trabalha com um modelo de dados baseado em comunidades e coleções, possibilitando aos usuários pesquisar e navegar nas publicações, através de ferramentas de buscas internas.

5.1 A estrutura de informação dos repositórios digitais.

As principais ferramentas que permitem a implantação de repositórios institucionais apresentam características semelhantes quanto à forma com que armazenam seus dados. Todas elas estão amparadas por uma estrutura que define um banco de dados, relacional em grande parte das vezes, para armazenar as informações que são postadas pelos mais variados tipos de usuários.

Dentro do contexto de armazenamento, o que se vê é a utilização de banco de dados relacionais, onde cada ferramenta implementa um

diferente modelo lógico de dados para que as informações sejam armazenadas.

De modo geral, os produtos de banco de dados mais utilizados pelas ferramentas são: Postgresql, Oracle e Mysql, não necessariamente nesta ordem.

Como neste trabalho a demonstração de aplicação será realizada com o uso do software DSPACE, utilizar-se-á o mesmo como exemplo para apresentação das características estruturais de um repositório digital.

O Dspace oferece a possibilidade de ser implantado com o uso do Postgresql ou do Oracle, ficando a cargo da equipe de implantação a escolha da melhor opção, de acordo com o tipo de aplicação e da estrutura funcional da instituição que receberá o repositório.

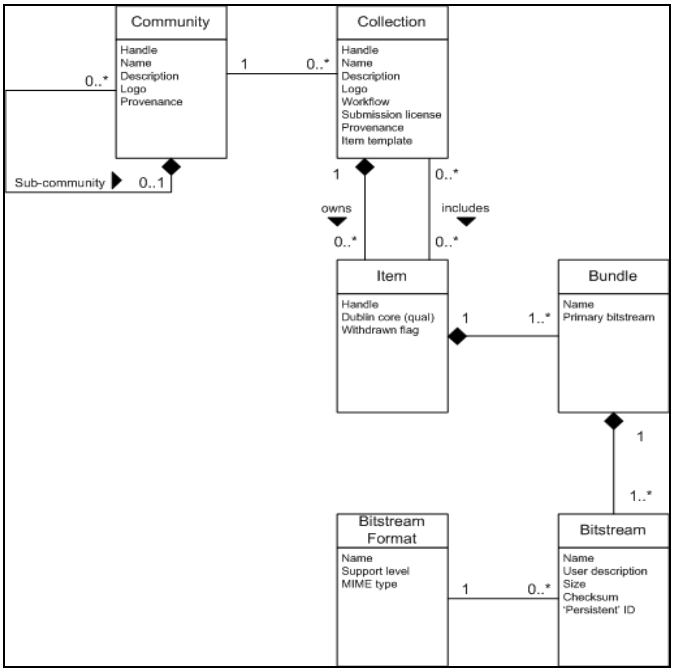


Figura 23 - Modelo Lógico de Banco de Dados – Dspace
Fonte: Documentação DSPACE

Através da figura 23, pode-se verificar o modelo lógico utilizado pelo Dspace e oferecido como referência em sua documentação. Ele é composto basicamente por seis entidades, sendo que cada uma representa um papel específico no armazenamento de informações:

- Community: Comunidade, como a Universidade Estadual Paulista ou o Departamento de Ciência da Informação.
- Collection: Coleção, para separar as informações por grupos, como, “relatórios técnicos”, artigos, material de aula.
- Item: Um relatório, um artigo, uma apresentação.
- Bundle: Grupo ou pacote de informações que representa um documento.
- Bistream: Informações específicas sobre os arquivos (documento, imagem, arquivo de dados) que compõem um recurso.
- Bistream format: Especificação do formato do arquivo que compõe o recurso, como PDF. TXT, DOC.

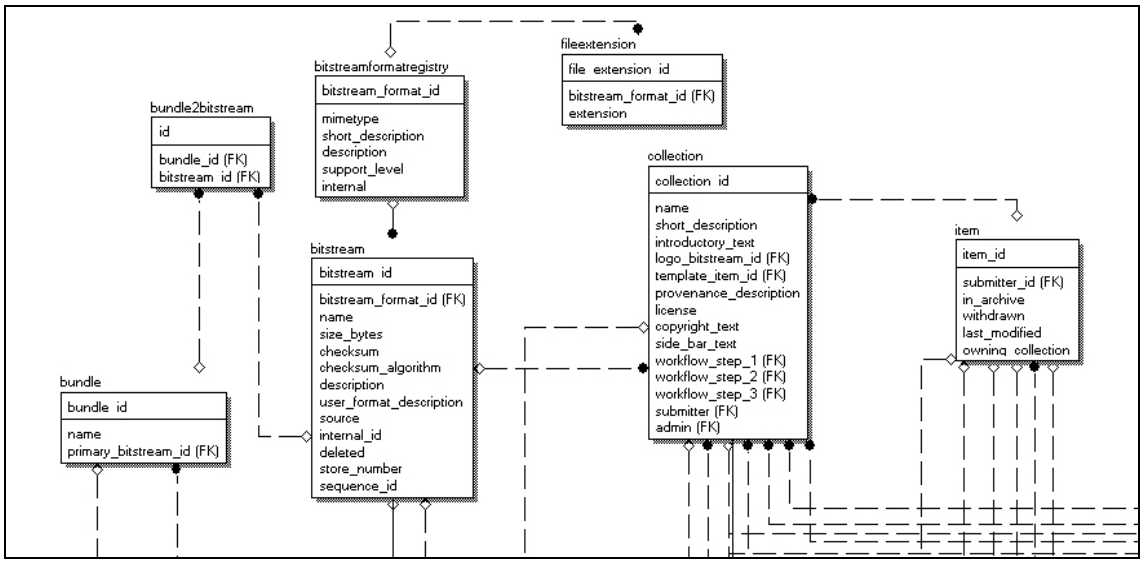


Figura 24 - Parte do Modelo Físico do Dspace.
Fonte: Documentação DSPACE

O modelo lógico apresentado, apesar de parecer simples, embute em suas informações um conjunto de outras informações que registram tudo o que os usuários precisam no momento de armazenar ou de recuperar informações em um repositório institucional, conforme o modelo sugerido pelas ferramentas.

O modelo físico, que é a representação real de implementação do banco de dados, apresenta um conjunto completo e rico em detalhes para que toda a estrutura de informação possa ser implementada e armazenada no banco de dados. Através da figura 24, é possível verificar que a maneira que as entidades são apresentadas no modelo lógico tem função apenas para efeito de entendimento do contexto global de informações.

O modo de armazenar fisicamente as informações dentro de um repositório não interessa aos usuários, em grande parte dos casos, ficando muito mais a cargo da equipe de desenvolvimento ou atualizações da ferramenta, porém essa estrutura física está baseada em um conjunto mínimo de informações que deve ser seguido, para que as informações armazenadas em repositórios possam ser interoperáveis.

A necessidade de interoperabilidade dos dados surgiu juntamente com o crescimento de iniciativas para resolver o problema da disseminação da informação, apresentada no começo deste capítulo, visto que, com a necessidade de desenvolver estruturas que permitissem o armazenamento e consequente recuperação da informação em repositórios institucionais, cada instituição iniciou o desenvolvimento do seu próprio modelo de informações.

Garantir a interoperabilidade e integração entre os inúmeros sistemas de informação é inquestionável. A criação de repositórios de dados e serviços comuns/partilhados exige a implantação de soluções que permitam a integração eficaz e segura entre diferentes sistemas. Assim, pode definir-se interoperabilidade como o processo através do qual se assegura que diferentes sistemas, procedimentos e a própria cultura de uma organização sejam maximizados, permitindo a recuperação e a utilização constante da informação (MILLER, 2000 in SAYÃO, 2007). O assegurar da interoperabilidade implica a reestruturação e remodelação dos procedimentos organizacionais, nomeadamente nas relações com os utilizadores e com o uso da informação. Nesse sentido, têm-se desenvolvido uma série de padrões e protocolos de comunicação, transferência,

armazenamento e codificação de informação, como o Z39.50, o OAI-PMH e o XML (SAYÃO, 2007).

Dada a divergência entre as estruturas utilizadas nas ferramentas utilizadas como repositório, foi escolhido um modelo básico de dados que permitia a troca de informações entre repositórios digitais, que é o OAI-PMH.

Segundo o site oficial do protocolo OAI-PMH (2004),

> O protocolo OAI-PMH é um mecanismo para transferência de dados entre repositórios digitais. É uma interface que um servidor de rede pode empregar para que os metadados de objetos residentes no servidor estejam disponíveis para aplicações externas que desejem coletar esses dados. Essa interface tem duas propriedades: interoperabilidade e extensibilidade. A interoperabilidade decorre da obrigatoriedade embutida no protocolo para implementação do padrão Dublin Core. Assim todos os repositórios que utilizam o protocolo OAI podem trocar metadados. Já a extensibilidade advém da oportunidade de se criar ou utilizar também padrões de metadados diferentes do Dublin Core. Descrições específicas para uma comunidade ou especificidade de metadados para satisfazer necessidades especiais podem ser criadas ou adaptadas de forma a funcionarem com o protocolo OAI.

O uso do OAI-PMH, que é baseado no padrão DC, oferece à ferramenta a estrutura necessária para que as informações sejam posteriormente interoperáveis.

Marcondes (2005, p.100) indica:

> a vantagem do uso do OAI-PMH consiste em permitir a coleta automática de metadados de documentos armazenados em arquivos de publicações eletrônicas os provedores de dados. Os metadados são coletados conforme o Dublin Core (padrão internacional), pois é mais específico para informação bibliográfica.

É inegável que as ferramentas que possibilitam a implementação de repositórios devem estar de acordo com a estrutura necessária para implementar o protocolo OAI-PMH, visto que ele facilita o processo de interoperabilidade e, consequentemente, a troca de informação entre repositórios e serviços.

É importante ressaltar que o protocolo OAI-PMH implementa apenas os elementos principais do padrão DC, ficando a extensibilidade restrita apenas ao repositório em que o material/documento está depositado.

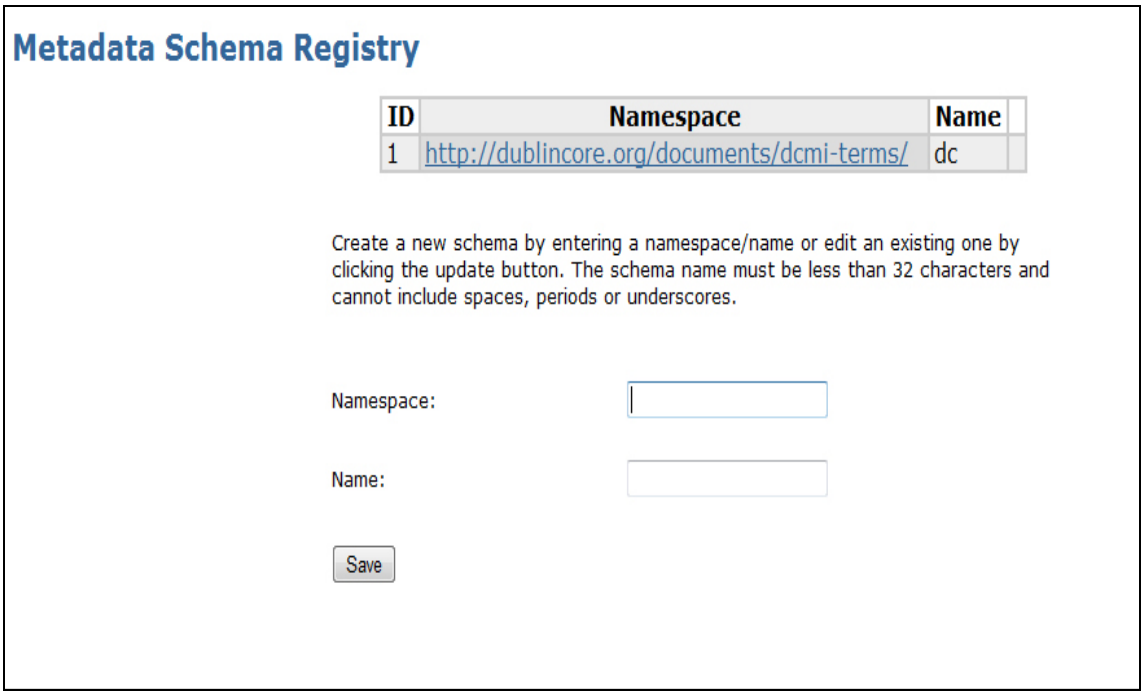


Figura 25 - Inserção de outro padrão de metadados na ferramenta Dspace. Área administrativa do software.

Fonte: Dspace

É possível verificar, na figura 25, que o Dspace já traz em sua estrutura original o padrão de metadados DC Qualificado, mas também dá liberdade para que os administradores da ferramenta cadastrem e sugiram outro tipo de padrão de metadados. O uso exclusivo de outros padrões de metadados impede o uso do protocolo OAI-PMH.

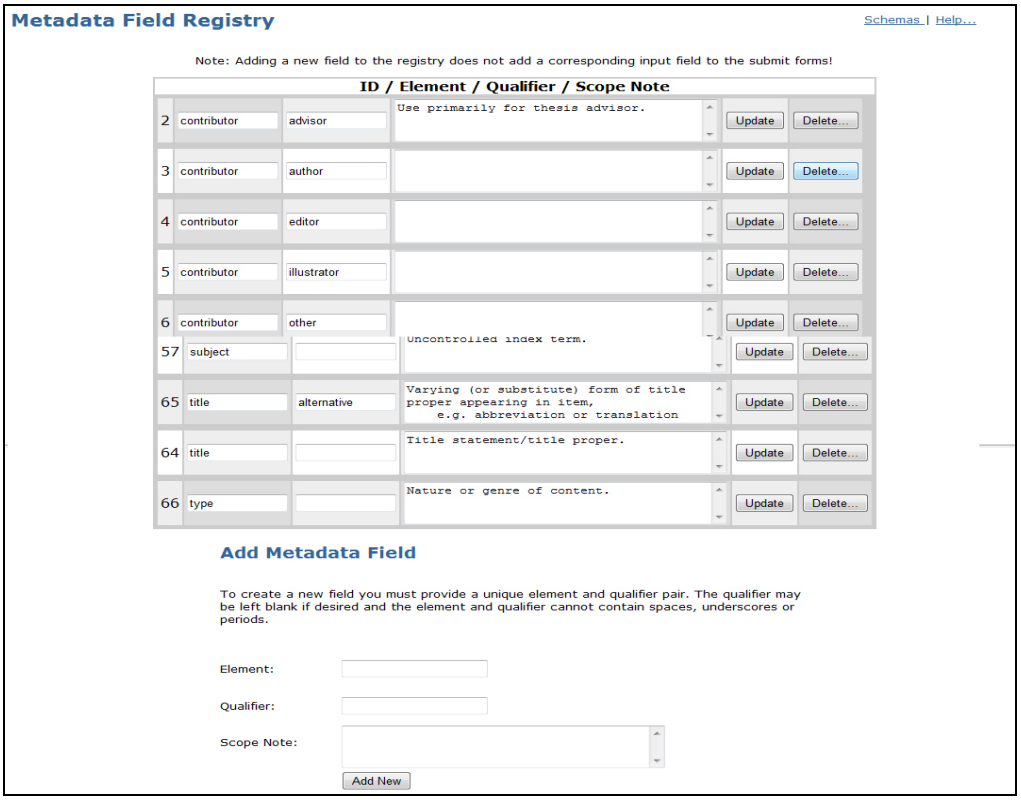


Figura 26 - Alteração do padrão DC Qualificado na ferramenta Dspace. Área administrativa do software.

Fonte: Dspace

A ferramenta Dspace também possibilita a extensibilidade do padrão DC qualificado (figura 26) que ela já embute originalmente em seu código e que conta, na versão 1.5.0, com 70 elementos. Dá a liberdade ao administrador do sistema de inserir novos elementos, além de excluir e alterar os que já existem.

Destaca-se que os softwares que implementam repositórios apresentam uma camada lógica, baseada em padrão de metadados e uma estrutura física que indica o uso de um banco de dados relacional. Esse tipo de estrutura funcional é muito claro quando se utilizam os padrões difundidos e empregados na Ciência da Informação, e sugere em alguns casos um repensar sobre grande parte da teoria de modelagem de dados utilizada no contexto do desenvolvimento de sistemas de informação.

Grande parte da estrutura apresentada em relação a padrões de metadados e de protocolos que permitem a troca de informações através de

um modelo de interoperabilidade sugerido tem como princípio fundamental possibilitar a troca de informações e a recuperação mais adequada de informações aos usuários.

5.2 A recuperação de informação em repositórios digitais.

A recuperação de informações em repositórios digitais apresenta um grande diferencial em relação à recuperação de informações na Web, pois parte de um princípio de que a informação foi registrada e armazenada de forma adequada, seguindo padrões de catalogação e uso de metadados e com conteúdo e estrutura de informação muito bem delimitada e separada, baseada em conceitos que se preocupam com a recuperação da informação, como o uso de estrutura e formatos de representação da informação previamente estudados.

Pelo contexto apresentado até este momento, é possível perceber que a recuperação de informação pode ser segmentada e se tornar específica para atender à busca em determinados campos que estão diretamente relacionados aos elementos do padrão de metadados utilizado.

Apesar de a estrutura de armazenamento sugerir um tipo de recuperação mais apropriado ao usuário, ela continua sendo feita de forma sintática, buscando, dentro do conjunto de informações armazenadas, palavras que tenham mesma grafia, e utilizando a técnica baseada no modelo booleano e na teoria de conjuntos, possibilitando apenas o cruzamento de elementos da estrutura na busca de informação.

A apresentação dos resultados também não sugere novidades em relação às principais ferramentas de busca encontradas na Web, tendo características limitadas e utilizando como principal formato a apresentação de uma lista de informações que remetem a um link, onde naturalmente está o recurso.

Outro ponto que pode ser abordado no contexto de recuperação é que o conjunto de informações disponíveis em repositórios digitais é muito grande e bem estruturado. Portanto, além da simples recuperação de informações baseadas em expressões dos usuários, poderiam ser apresentados cruzamentos de informações dentro do próprio contexto dos dados armazenados, com apresentação de rankings e possíveis relacionamentos entre objetos que têm o mesmo conteúdo, autor ou instituição, por exemplo. Isso poderia ser caracterizado dentro de uma estrutura no formato de redes, permitindo relacionar informações que não têm relação sintática, mas sim semântica ou de associação por alguma outra característica.

De modo geral, a recuperação da informação em repositórios pode ser muito explorada e evidentemente melhorada, dadas as características estruturais pelas quais estes objetos digitais são constituídos.

No capítulo 7 será feita uma abordagem específica para a recuperação da informação, já baseada no modelo proposto nesta pesquisa.

5.3 Os recursos e funcionalidades da Web 2.0 em repositórios digitais

A Web 2.0 é caracterizada pela implementação de itens de tecnologia e também pela construção da inteligência coletiva através do desenvolvimento de um tema.

De modo geral, os repositórios têm o perfil de permitir a uma comunidade a disponibilização do seu material, através de uma plataforma acessível via web, portanto é possível aplicar praticamente todos os recursos da Web 2.0 em repositórios digitais.

Dentro deste contexto, verificaram-se os itens oferecidos pela estrutura básica dos principais softwares, em suas versões mais atuais, além de um grande número de repositórios instanciados com as

ferramentas, para estruturação de repositórios digitais: Dspace[29], E-prints[30], Fedora[31]. A escolha dos três softwares foi motivada pela quantidade de repositórios atualmente implementados com eles, por manterem equipes de desenvolvimento trabalhando na evolução e atualização das versões e porque são oferecidos sob licença open-source, dando liberdade à instituição ou à equipe de programação para ampliar os recursos oferecidos.

Além dos softwares que exigem licença de uso, não foram observados repositórios construídos sobre uma plataforma proprietária, ou seja, desenvolvida por uma equipe técnica específica de uma instituição, unicamente para aquele repositório.

O quadro 3 apresenta o resultado da abordagem:

	Dspace 1.5	E-print 3	Fedora 3.2
RSS	S	S	S
Tag Clouds	N	N	N
Mashup	N	N	N
Interfaces Ricas	N	N	N
Comentários	S - Add	N	N
Blog	N	N	N

QUADRO 3 - RELAÇÃO ENTRE SOFTWARE REPOSITÓRIOS X RECURSOS WEB 2.0

Os três softwares para implementação de repositórios digitais analisados apresentam características muito semelhantes, principalmente em relação aos recursos de Web 2.0 nele implementados.

Foi verificado que o único recurso disponível nos três softwares foi RSS. O recurso de RSS é realmente o mais simples no contexto de desenvolvimento técnico, e, portanto, mais disponível. Talvez por isso esteja

[29] http://www.dspace.org
[30] http://www.eprints.org
[31] http://fedora-commons.org/

presente em todos eles. Todos os softwares apresentam opções de RSS nas versões 1.0 e 2.0.

Recursos como Tag Clouds, Mashups e ferramenta para Blog não estão disponíveis em nenhum deles. Como as ferramentas são Open-Source, foi realizada uma verificação em aproximadamente 80 repositórios digitais que usam as ferramentas citadas e nenhum deles recebeu alteração de estrutura para que os recursos fossem implementados.

O software Dspace, ao contrário dos outros dois (E-print e Fedora), apresenta, em sua versão ampliada com o uso da ferramenta Manakin, responsável por melhorar a interface de apresentação do Dspace, um modelo de apresentação diferenciado, facilitando, ao responsável pela implementação do repositório, o uso de Interfaces Ricas, porém somente o uso da ferramenta não apresenta recursos suficientes que possam ser caracterizados como interfaces ricas.

Apesar de não citado anteriormente como um dos recursos que caracterizam a Web 2.0, os comentários em postagens são um recurso que tem aparecido constantemente nos blogs e em portais de notícias, para que os usuários possam de certa forma interagir com o conteúdo postado. Como esse recurso facilita o processo de Inteligência Coletiva, não construção, mas pela possibilidade de interação, foi incluído como recurso observado nas ferramentas de repositórios.

Nenhum dos softwares verificados apresenta o recurso de comentários, porém o Dspace apresenta em sua página de Add-nos e Extensions, que são módulos do programa desenvolvidos por terceiros colaboradores, um Add-on desenvolvido pela Universidade do Minho [32]de Portugal, que possibilita a implementação deste recurso ao repositório que utiliza as versões mais recentes (acima de 1.4.2) do Dspace.

Apesar de os recursos de Web 2.0 já fazerem parte da maioria dos portais mais populares na Web, eles ainda são restritos e pouco utilizados em ferramentas que implementam repositórios digitais.

5.4 Os recursos e funcionalidades da Web 3.0 em repositórios digitais

Os repositórios digitais são estruturas de informação recentes, portanto já contemplam alguns dos principais recursos sugeridos como parte da estrutura para se constituir uma Web Semântica.

Apesar de o contexto da Web Semântica estar baseado em ambientes abertos, é possível pensar que as tecnologias apontadas para o desenvolvimento da Web 3.0 possam ser aplicadas em ambientes estruturados, com o objetivo de desenvolver uma estrutura de recuperação da informação baseada em conteúdos e, com auxílio de uma ontologia, criar um mapa de relação quando no momento da busca.

Os principais softwares, já indicados neste capítulo, para implementação de repositórios utilizam uma estrutura que propicia o emprego de tecnologias da Web 3.0 dentro de seu contexto, apesar de não estarem preparados e muito menos direcionados para este conceito.

Em todos os três softwares — Dspace, E-prints e Fedora — existe o uso de um banco de dados relacional para que os dados sejam armazenados. A estrutura em que as informações são armazenadas segue critérios diferentes, porém conta com boa alternativa de estrutura de informações quanto ao uso de modelos em banco de dados relacionais.

O fato de os softwares primarem pela interoperabilidade e disporem de estrutura informacional para trocar informações através do protocolo OAI-PMH indica o uso de estrutura de informação em formato XML, fator que contribui para a implementação de técnicas e métodos de Web 3.0 nos repositórios.

Entre as principais características que os softwares apresentam, que possibilitam efetivar o uso dos padrões da Web 3.0 em repositórios, está a estruturação de informações através do uso dos metadados. A opção pelo uso do padrão DC, reconhecido e recomendado pelo W3C, facilita a integração de outras tecnologias ao conteúdo dos repositórios.

---

[32] http://www.uminho.pt

A oferta de uma estrutura que pode contribuir para o desenvolvimento de busca semântica nos repositórios é certa, porém o único software que implementa, através de add-on, o uso de ontologias para publicação das palavras chaves é o Dspace, incluindo o uso da linguagem OWL para descrever a ontologia. Apesar de o Dspace oferecer a possibilidade de estruturas de representação do conhecimento, como vocabulários controlados ou ontologias, para descrição das palavras-chave, os outros dois softwares analisados não fazem qualquer referência a este conceito, e também a nenhum outro que possa dar o entendimento de que há uma convergência para recuperação semântica nestes ambientes.

A utilização de ontologia OWL é um grande passo para construir busca semântica e aplicar relacionamento entre os termos através do uso das classes disponibilizadas nas tecnologias.

Dessa forma, fica claro que os repositórios são ambientes informacionais que, se adaptados, podem efetivamente melhorar muito o processo de descrição e, consequentemente, de recuperação da informação, porque o formato estrutural em que foram construídos é totalmente propício à utilização dos conceitos de Web Semântica.

No próximo capítulo será apresentado o modelo “Representação Iterativa” a ser aplicado em repositórios digitais científicos, a fim de aproveitar essa pré-disposição funcional dos repositórios e de torná-los modelo de recuperação semântica de informação.

6 REPRESENTAÇÃO ITERATIVA, MODELO DE ESTRUTURA PARA DESCRIÇÃO, ARMAZENAMENTO, REPRESENTAÇÃO DE RECURSOS E RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS CIENTÍFICOS

Os capítulos anteriores apresentaram os conceitos básicos e necessários para o entendimento da "Representação Iterativa" proposta nesta tese, possibilitando estabelecer argumentos para o desenvolvimento desse modelo para repositórios digitais.

Este capítulo apresenta: o modelo "Representação Iterativa", que deve transformar um repositório digital científico em uma ferramenta apta a descrever, armazenar e recuperar informação, permitindo a recuperação semântica e a construção coletiva de uma estrutura relacional semântica de informações através de Folksonomia Assistida; e as técnicas utilizadas no desenvolvimento da estrutura sugerida.

Ressalte-se que a ferramenta Dspace servirá como apoio nos momentos em que for necessário criar relações do modelo com uma ferramenta real, além de expressar, através de exemplos, fórmulas ou construções conceituais.

Iniciar-se-á com uma abordagem sobre a estrutura funcional do Dspace em relação a sua camada de metadados, apresentando de forma objetiva a relação existente entre o Dublin Core e a modelagem de banco de dados desenhada para armazenar as informações.

Em seguida, será apresentado o estudo de Catarino (2009), que prevê a utilização de Folksonomia em repositórios digitais, visto que o estudo realizado pela autora será importante para compor a estrutura de funcionamento do modelo de Representação Iterativa.

Finalmente será descrita estrutura da Representação Iterativa, e, em seguida, a descrição de cada fase do processo de construção da informação quando da utilização do modelo proposto.

6.1 Armazenamento – a relação entre Dublin Core e Banco de Dados

No capítulo anterior, quando se tratou de repositórios, apresentou-se o modelo lógico e parte do modelo físico de banco de dados da ferramenta Dspace, modelos que garantem o armazenamento de informações que são registradas em um repositório digital.

Importante ressaltar que, diferente de um simples registro de banco de dados ou então de livre armazenamento de um documento, o processo de autoarquivamento de objetos digitais em um repositório digital científico é um pouco mais complexo e exige dedicação do usuário, que deverá descrever a informação de maneira coesa ao autoarquivar seu objeto digital.

A atividade de inserir informações em um repositório digital compreende o processo de inicialmente descrever o conjunto de informações que representa os metadados do objeto a ser inserido e, na sequência, realizar o envio do arquivo principal e também dos arquivos complementares, se houverem, para que todo o conjunto de informações seja armazenado no repositório.

Dá-se o nome de arquivos binários ao arquivo principal e seus complementares que podem estar na forma de documentos, planilhas, imagens, audios, vídeos, etc. Estes arquivos são inseridos de forma a ficarem armazenados no servidor em um conjunto de diretórios pré-estabelecidos pela ferramenta.

O armazenamento interno de informações se dá de duas maneiras que completam o processo: através da gravação dos metadados, em uma estrutura de banco de dados, de forma estruturada; e também através do armazenamento do arquivo full-text, de forma não estruturada, através de um ou mais arquivos binários.

O armazenamento de informações estruturadas guarda no banco de dados do repositório as informações pertinentes aos metadados que foram descritos pelo usuário, assim como as informações complementares a

respeito da comunidade e coleção de que o objeto faz parte. Informações a respeito dos arquivos binários, como tamanho, tipo de arquivo e nome, também são armazenadas no banco de dados.

Para armazenar as informações sobre as comunidades e coleções, e a relação de qual coleção faz parte de qual comunidade, o Dspace define, respectivamente, três tabelas físicas denominadas: community, collection e community2collection (figura 27).

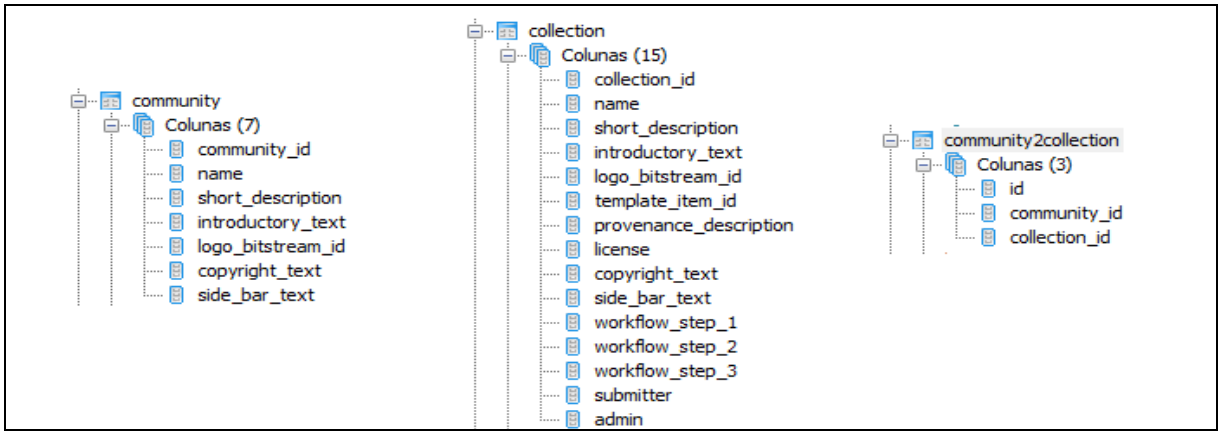


Figura 27 – Tabelas community, collection e community2collection

Fonte: Dspace

A tabela community é utilizada para armazenar as comunidades que fazem parte do repositório; a tabela collection armazena as informacões a respeito das coleções; e a tabela community2collection tem a função de armazenar o relacionamento entre as comunidades e coleções, ou seja, as coleções que fazem parte de cada uma das comunidades.

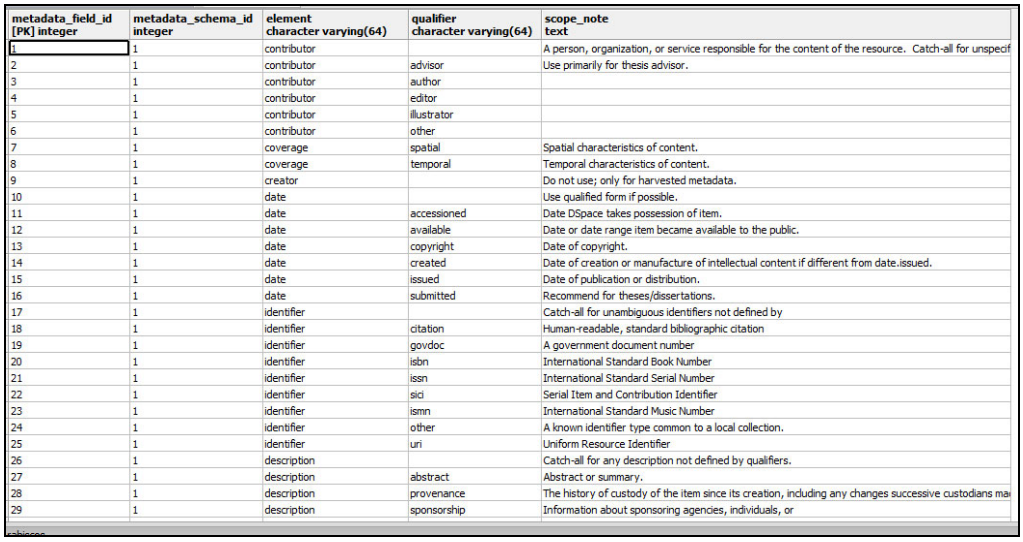
As comunidades e coleções são definidas, organizadas e gerenciadas pelos administradores, de forma que ofereçam ao usuário uma organização lógica a respeito do domínio de conhecimento em que o repositório está inserido.

Para conceber o armazenamento interno das informações relativas aos objetos que estão sendo depositados no repositório, em sua grande parte por pesquisadores, o banco de dados define um conjunto de tabelas que deve armazenar desde a informação do próprio usuário que está fazendo o

depósito, incluindo data, até o conjunto de informações que compõe os metadados do recurso a ser depositado. No Dspace, as tabelas físicas responsáveis por armazenar as informações do objeto digital depositado são:

- item: responsável por armazenar as informações sobre o usuário que fez o depósito, e definir um número único para o objeto, além da data em que foi realizada a última alteração no objeto;
- collection2item: armazena a informação referente a qual coleção pertence o recurso que está sendo inserido (estabelece o relacionamento);
- metadatavalue: armazena as informacões dos metadados do recurso que está sendo inserido. Essa tabela faz uma ligação direta com outras duas tabelas: metadataschemaregistry e metadatafieldregistry.

Assim como outras ferramentas, o Dspace permite o cadastro de mais de um esquema (formato) de metadados para ser utilizado, portanto, além do já pré-definido Dublin Core, podem-se cadastrar outros formatos de metadados que já foram desenvolvidos e definidos por alguma comunidade específica. Isso implica que a ferramenta não está restrita a apenas um formato de metadados. A tabela metadataschemaregistry é a responsável por registrar cada um dos esquemas de metadados que o repositório suporta, portanto cada registro da tabela representa um esquema de metadados diferente.



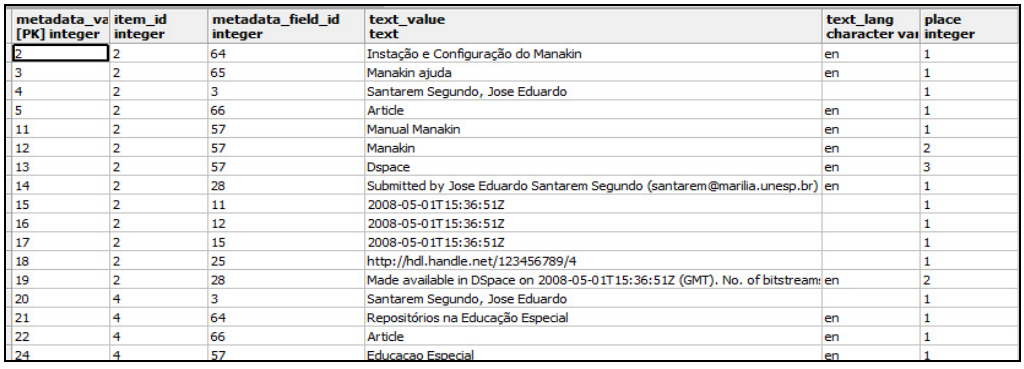
metadata_field_id [PK] integer	metadata_schema_id integer	element character varying(64)	qualifier character varying(64)	scope_note text
1	1	contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecif
2	1	contributor	advisor	Use primarily for thesis advisor.
3	1	contributor	author	
4	1	contributor	editor	
5	1	contributor	illustrator	
6	1	contributor	other	
7	1	coverage	spatial	Spatial characteristics of content.
8	1	coverage	temporal	Temporal characteristics of content.
9	1	creator		Do not use; only for harvested metadata.
10	1	date		Use qualified form if possible.
11	1	date	accessioned	Date DSpace takes possession of item.
12	1	date	available	Date or date range item became available to the public.
13	1	date	copyright	Date of copyright.
14	1	date	created	Date of creation or manufacture of intellectual content if different from date.issued.
15	1	date	issued	Date of publication or distribution.
16	1	date	submitted	Recommend for theses/dissertations.
17	1	identifier		Catch-all for unambiguous identifiers not defined by
18	1	identifier	citation	Human-readable, standard bibliographic citation
19	1	identifier	govdoc	A government document number
20	1	identifier	isbn	International Standard Book Number
21	1	identifier	issn	International Standard Serial Number
22	1	identifier	sici	Serial Item and Contribution Identifier
23	1	identifier	ismn	International Standard Music Number
24	1	identifier	other	A known identifier type common to a local collection.
25	1	identifier	uri	Uniform Resource Identifier
26	1	description		Catch-all for any description not defined by qualifiers.
27	1	description	abstract	Abstract or summary.
28	1	description	provenance	The history of custody of the item since its creation, including any changes successive custodians ma
29	1	description	sponsorship	Information about sponsoring agencies, individuals, or

Figura 28 – Tabela metadatafieldregistry (Dspace)

Fonte: Dspace

Outra tabela física citada, metadatafieldregistry, armazena os itens (elementos) referentes aos esquemas de metadados registrados no repositório, ou seja, cada registro da tabela representa um elemento de um dos esquemas de metadados (figura 28). Os campos da tabela representam: o esquema de metadados a que o elemento faz parte (metadata_schema_id), o nome do elemento (element), o qualificador do elemento (qualifier), possibilitando o registro de elementos qualificados, conforme visto no tópico sobre Dublin Core (capítulo 3), e ainda um último campo que permite a gravação de um texto de descrição sobre o elemento (scope_note).

No exemplo apresentado na figura 28, é possível verificar parte do esquema que acompanha o Dspace em sua instalação.



metadata_va [PK] integer	item_id integer	metadata_field_id integer	text_value text	text_lang character var	place integer
2	2	64	Instação e Configuração do Manakin	en	1
3	2	65	Manakin ajuda	en	1
4	2	3	Santarem Segundo, Jose Eduardo		1
5	2	66	Article	en	1
11	2	57	Manual Manakin	en	1
12	2	57	Manakin	en	2
13	2	57	Dspace	en	3
14	2	28	Submitted by Jose Eduardo Santarem Segundo (santarem@marilia.unesp.br)	en	1
15	2	11	2008-05-01T15:36:51Z		1
16	2	12	2008-05-01T15:36:51Z		1
17	2	15	2008-05-01T15:36:51Z		1
18	2	25	http://hdl.handle.net/123456789/4		1
19	2	28	Made available in DSpace on 2008-05-01T15:36:51Z (GMT). No. of bitstream	en	2
20	4	3	Santarem Segundo, Jose Eduardo		1
21	4	64	Repositórios na Educação Especial	en	1
22	4	66	Article	en	1
24	4	57	Educação Especial	en	1

Figura 29 – Tabela metadatavalue – Dspace

Fonte: Dspace

Os metadados do objeto a ser inserido no repositório digital, descritos pelos usuários, são armazenados na tabela metadatavalue (figura 29). Nesta tabela, cada informação armazenada deve estar devidamente relacionada a tabela item, através do campo item_id, de forma que o registro represente uma informação de um determinado item. Dessa forma, verifica-se que o campo item_id da tabela (figura 29) apresenta nos primeiros registros o número 2 e nos ultimos registros o número 4, definindo que os primeiros registros são de um item com código 2 e os outros restantes são de um outro item com código 4, ou seja, todos as informações apresentadas na figura 28 são parte de dois depósitos diferentes.

É possível observar também, na figura 29, que o campo metada_field_id faz relação ao elemento de metadado registrado na tabela metadatafieldregistry, apresentada através da figura 28. Assim, verifica-se que, neste exemplo, o terceiro registro armazenado na tabela metadavalue representa a informação sobre o elemento contributor, com qualifier author. A informação sobre o valor que deverá ser registrado para esse elemento está no campo text_value.

Através deste conjunto de relações, construídas por um modelo relacional, o Dspace armazena as informações necessárias para guardar um objeto depositado em um repositório digital.

Além das tabelas apresentadas, outras tabelas do modelo físico também são utilizadas para armazenar detalhes de parte do conjunto de informações do depósito, porém, dado o foco deste trabalho, o conjunto de informações apresentadas será suficiente para o entendimento do modelo de Representação Iterativa.

Nesta pesquisa, o foco não está em detalhar a estrutura de banco de dados do Dspace, nem tampouco de qualquer outra ferramenta para repositórios, porém é importante apresentar uma parte da estrutura que é responsável pelo armazenamento de informações e a maneira como a

ferramenta gerencia esse conjunto de informações em sua estrutura relacional de banco de dados, visto que, posteriormente, na construção do modelo proposto nesta pesquisa, deverá ser ampliado o modelo de banco de dados da ferramenta.

6.2 Folksonomia em repositórios digitais científicos

No capítulo 3 já foram abordados o termo e o conceito de Folksonomia, e ainda verificou-se, no capítulo 5, que essa funcionalidade não está disponível nas principais ferramentas de repositórios digitais disponíveis para implantação e uso.

No modelo Representação Iterativa, proposto nesta tese, considera-se a Folksonomia como funcionalidade fundamental, pois caracteriza a construção da informação de forma coletiva e prioriza a participação do usuário, em grande parte pesquisadores, na construção do vocabulário do domínio de conhecimento em que o repositório está inserido.

Dentro deste contexto, é importante ressaltar o trabalho de Catarino (2009), que aborda de forma direta o uso de Folksonomia em repositórios digitais.

Segundo Catarino (2009, p. 59),

> Pressupõe-se que a folksonomia permite uma nova forma de organização de recursos da Web e que, naturalmente, poderá também ser adoptada pelos Repositórios Institucionais para que seus utilizadores tenham uma forma de organizar os recursos conforme suas necessidades.
>
> Além de servir como uma forma de organização individual, julga-se que as etiquetas atribuídas pelos utilizadores possam ser aproveitadas pelos gestores dos Repositórios para enriquecer a informação relativa aos recursos neles depositados. As etiquetas podem ser relacionadas com propriedades do DC e outras propriedades complementares, enriquecendo, assim, a organização dos recursos sem comprometer a interoperabilidade dos seus metadados.

Através de sua pesquisa, a autora verificou que as tags utilizadas pelos usuários em ambientes folksonômicos podem ser representadas em grande parte por elementos do padrão Dublin Core, por isso baseou-se em coleta de informações nos sites Delicious e Conotea.

No cômputo geral, os dados representavam 50 recursos, etiquetados por 15.381 utilizadores, com 5.098 etiquetas atribuídas. Considerando que uma etiqueta podia ser atribuída a vários recursos e por vários utilizadores, optou-se por registar o total de ocorrências das etiquetas 79.146 (CATARINO, 2009).

Para garantir uma identificação segura, o processo de organização das etiquetas necessitou de alguns ajustes, conforme relata Catarino (2009, p. 94):

> As etiquetas analisadas foram agrupadas em suas formas variantes (singular/plural, maiúsculas/minúsculas, idiomas, grafia, siglas e abreviaturas). Este procedimento foi realizado para facilitar posteriormente a identificação das propriedades. Pressupôs-se que o agrupamento das etiquetas facilitaria a compreensão das mesmas e consequentemente a identificação das propriedades. Como resultado deste agrupamento, pode-se perceber melhor o significado e agilizar o processo de identificação das propriedades.

Os resultados alcançados demonstram que grande parte das tags que foram inseridas pelos usuários são relativas à descrição do assunto, caracterizada pelo elemento subject do padrão de metadados Dublin Core.

Verificou-se, portanto, que a propriedade Subject podia ser relacionada com 52,9% do total geral de ocorrência de Key-tags e a 87,3% da ocorrência de Key-tags relacionadas com elementos do DC (CATARINO, 2009).

Este contexto, devidamente estudado por Catarino, permite verificar que grande parte das informações sugeridas através de tags é utilizada com relação ao campo assunto, do documento ou link que este deverá indexar.

Catarino (2009, p. 149) sugere a alteração do esquema de metadados Dublin Core, no contexto de repositórios institucionais, para que os mesmos possam receber a funcionalidade de Folksonomia.

> O *Social Tagging Application Profile* (STAP) foi criado para declarar termos de metadados que são propriedades complementares às já existentes no DC para a descrição de recursos de repositórios institucionais que implementem funcionalidades de *social tagging* ou importem etiquetas de outros sistemas. Portanto, foi proposto para ser utilizado pelos repositórios institucionais que possuam uma folksonomia resultante das etiquetas atribuídas pelos próprios utilizadores dos recursos. A intenção é acrescentar valor à descrição tradicional permitindo que os próprios utilizadores registem os valores relativos às propriedades que descrevem o recurso. Pressupõem-se que desta forma serão ampliadas as possibilidades de organização e recuperação da informação de forma diferenciada.

Apesar da citação anterior, o modelo de Representação Iterativa tem preferência pela utilização do recurso de tag, indicando relação com o campo assunto, unicamente. Essa preferência é justificada pelos números de Catarino, que apresenta dados consistentes de que o campo assunto é realmente o mais utilizado para relacionamento das tags descritas com os documentos inseridos.

6.3 Representação Iterativa, estruturando o modelo

Conforme o trabalho vem sendo direcionado, é possível notar que a pesquisa sugere a construção de um modelo estrutural para repositórios digitais científicos, de forma que esses ambientes possam agregar funcionalidades que atuem no sentido de garantir ao usuário uma melhor interface de comunicação com o sistema e ainda evoluir no processo de recuperação da informação, possibilitando a apresentação de resultados baseados em relação semântica, baseada em associação de conteúdos, e não apenas em comparação sintática, como é realizado atualmente.

O modelo — Representação Iterativa — parte do princípio de que o usuário deverá ter uma interface diferente para inserção de dados no

repositório digital. A princípio, a única alteração em relação à interface padrão de descrição do recurso será no momento de informar as palavras-chave, visto que estes campos deverão vir com uma informação de que, além de configurar como palavras-chave, os dados descritos ali serão também utilizados como tags.

O fato de caracterizar o uso de tags já cria neste ambiente a ideia de que o ambiente tratará as palavras-chave como parte da concepção de Folksonomia, e, portanto, deverá implementar recursos que permitam a recuperação da informação em novos formatos, como uma nuvem de tags, por exemplo.

No momento em que o usuário iniciar o processo de descrição da tag deverá ocorrer uma intervenção do sistema, de forma que se caracterize um processo que se denomina Folksonomia Assistida.

6.3.1 Folksonomia Assistida, enriquecendo a descrição do recurso

Folksonomia Assistida é um processo de apoio ao usuário, no momento de definir os termos mais adequados para as tags que referenciarão seu trabalho depositado em um repositório digital. O processo é composto por duas partes principais.

A primeira parte implica que, para a implementação da Folksonomia Assistida, deverá ser alterada a interface de comunicação do usuário com o repositório, ou então desenvolvida uma nova interface, para a inserção de informações no campo palavra-chave, utilizado como referência para a inserção de conteúdo para as tags.

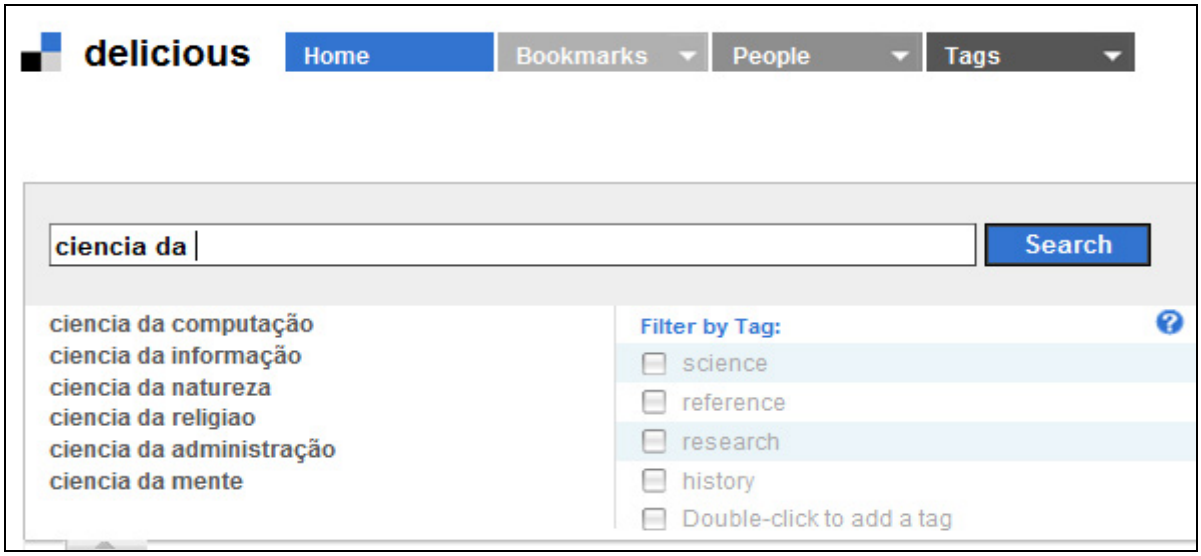


Figura 30 – Busca no Del.icio.us
Fonte: http://www.delicious.com

Nesse primeiro passo, deve-se apresentar ao usuário, no momento da digitação da tag, de forma sistemática, um conjunto de informações já previamente inseridas no sistema, como uma sugestão de tags. A busca de informação para fazer a sugestão é baseada em busca sintática.

Essa maneira de inserção de dados já é realizada no site Delicious (figura 30), e tem como característica a apresentação de sugestões conforme o usuário vai digitando o termo a ser registrado como tag. Tecnicamente, essa funcionalidade da Web 2.0, de interação com o usuário de forma rápida e sem recarregamento da página, são as já citadas interfaces ricas. Conforme já apresentado no capítulo 3, o recurso de "sugestão", utilizado pelo Delícious e também na concepção da Folksonomia Assistida, foi inicialmente apresentado pelo Google em sua ferramenta de busca, porém neste contexto tem sido adaptado para facilitar o processo de descrição do recurso pelo usuário.

O conjunto de informações que deverá ser apresentado ao usuário no momento que este estiver digitando será baseado nas tags já inseridas no sistema e também nos termos que fazem parte de uma estrutura de representação do conhecimento das áreas de especialidades que deverá estar associada ao repositório como parte do modelo estrutural proposto.

Assim que o usuário descrever as tags, aceitando ou não as sugestões, o sistema receberá a informação e dará início a um segundo passo para a concepção da Folksonomia Assistida.

No segundo passo, o repositório deverá receber os termos enumerados pelo usuário e proceder à pesquisa de relacionamento da informação dada pelo usuário em relação ao conjunto de informações internas que a ferramenta dispõe.

O processo de relacionamento em questão é justamente uma busca de relações dentro de uma estrutura de representação do conhecimento das áreas de especialidades, visto que esta pode ser caracterizada por um tesauro ou ainda por uma ontologia, que são instrumentos que permitem uma busca hierárquica horizontal, mas, principalmente, uma busca hierárquica vertical de relacionamento de termos.

Neste modelo, sugere-se o uso de uma estrutura de representação do conhecimento das áreas de especialidades, em qualquer um de seus instrumentos, porém no capítulo 4 foi abordado que a utilização de ontologias através da linguagem OWL permite agregar recursos e facilitar o processo de recuperação da informação, principalmente por ser uma linguagem que vem sendo aprimorada constantemente, e conta com indicação de uso pelo W3C.

A busca por termos relacionados em uma ontologia escrita com a linguagem OWL pode ser realizada através da linguagem Sparql, que tem como princípio justamente recuperar informações relacionadas em uma linguagem para descrição de ontologias.

Esse segundo passo da Folksonomia Assistida, além de recuperar termos relacionados em uma estrutura de representação do conhecimento das áreas de especialidades, deverá também buscar informações no conjunto de tags já inseridas no sistema, principalmente em seus relacionamentos horizontais. A busca por termos na estrutura de representação do conhecimento deverá acontecer em níveis pré-estabelecidos pelo administrador do ambiente, e a busca por relacionamentos horizontais no

conjunto de tags já descritas também poderá ser mediada pelo administrador, que deverá informar a quantidade de termos oferecidos para cada termo digitado pelo usuário. Esses conceitos poderão ser previamente parametrizados e adaptados conforme o repositório for sendo ampliado com novos depósitos.

A seguir, após essa busca interna por relacionamentos em relação ao termo descrito pelo usuário, o sistema apresentará novamente ao usuário um conjunto de termos que poderão ser aceitos de forma total ou parcial, ou ainda descartados pelo usuário, como sugestão final de tags para o recurso a ser inserido. Em todo esse processo, cabe ao usuário decidir as tags que melhor representem seu recurso digital dentro do domínio do repositório digital científico em que está sendo realizado o depósito.

A utilização de termos de uma estrutura de representação do conhecimento e também de tags já inseridas no sistema não tem o objetivo de engessar a criatividade do usuário, nem tampouco de descaracterizar o termo Folksonomia, pois o sistema permite claramente que o usuário decida livremente os termos que deverão ser utilizados como tags. A Folksonomia Assistida tem como principal característica oferecer ao usuário um conjunto de termos que já estão sendo empregados no sistema, de forma que ele possa usar a base de conhecimento do próprio repositório para qualificar a descrição de seu recurso.

A Folksonomia Assistida prima pela consistência das tags, de forma que o usuário do sistema evite abreviações, plurais/singulares ou ainda palavras que possam dificultar a recuperação da informação, posteriormente.

O processo de gravação das informações é efetivado quando o usuário definitivamente escolhe os termos que gostaria de usar como tags e grava as informações.

Ao decretar definitivamente o conjunto de dados que descrevem o objeto digital, o sistema receberá e armazenará no banco de dados o

conjunto de informações que o usuário escolheu para descrever o objeto digital.

O processo denominado Folksonomia Assistida, vem de encontro a necessidade fazer com que as tags tenham um grau maior de significado em relação ao objeto depositado, principalmente dentro do contexto em que está sendo utilizada.

Guy e Tonkin (2006, p. 1) afirmam que,

> Começamos por olhar para a questão das "tags malfeitas", um problema para o qual os críticos da Folksonomia fazem questão de aludir, e perguntar à comunidade que pesquisa sobre Folksonomia se há maneiras de compensar esses problemas [...]
>
> [...]Provavelmente, a grande falha dos sistemas de folksonomia atuais, é que os termos de marcação utilizados nesses sistemas são imprecisos. Os usuários dos sistemas que utilizam Folksonomia inserem livremente as tags, o que significa que as tags são muitas vezes ambíguas, excessivamente personalizadas e inexatas.

O uso da Folksonomia Assistida busca justamente melhorar a eficiência do uso de tags, permitindo ao usuário uma descrição livre para os objetos digitais a que deposita, porém de forma que possa se amparar no próprio conhecimento ja disponível no ambiente em que está utilizando.

6.3.2 Armazenando as tags de forma estruturada

A Representação Iterativa tem como princípio armazenar as tags definidas pelo usuário, portanto deve-se realizar uma alteração na estrutura de banco de dados que as ferramentas utilizam, criando um conjunto de tabelas que possa estabelecer o armazenamento e relacionamento dessas informações.

Nesse modelo se estabelece que uma nova tag, sempre que for inserida no repositório, deverá ser cadastrada em uma tabela; porém se a tag já existir no banco de dados, cadastrada em depósito anterior, apenas será

atribuído um incremento, no banco de dados, em relação à quantidade de vezes que a tag foi utilizada.

Outra característica da Representação Iterativa, baseada no Folksonomia, é efetivamente criar um relacionamento horizontal entre tags que descrevam o mesmo objeto digital. Esse processo constitui uma relação semântica entre os termos individualmente citados, e, dentro de um contexto de domínio do conhecimento restrito, estabelece uma relação entre termos, de forma que possam ser recuperados posteriormente.

Cada vez que um conjunto de termos for inserido, é estabelecida a relação, e assim vai se fortalecendo a estrutura de ligação entre as tags. Portanto, toda vez que houver um mesmo relacionamento entre termos, deverá apenas ser incrementada a quantidade de vezes que o relacionamento acontece, sem a necessidade de se recadastrar a informação no banco de dados.

Essa estrutura de informação que relaciona termos permite que se crie um grafo de tags, onde cada tag será representada por um vértice e a quantidade de relações entre as tags será representada visualmente pela largura, considerando o peso, da aresta que liga os vértices.

O armazenamento das tags nesse formato de relação horizontal permite constituir uma rede de informações.

As redes são consideradas um dos novos fenômenos de estudo na Ciência da Informação, e a relação das tags no formato de redes permite estudos mais aprofundados posteriormente do conteúdo que será gerado pelo repositório digital científico que implementar o modelo de Representação Iterativa.

Segundo Matheus e Silva (2009, p. 243),

> Na análise de redes o foco do estudo é nos relacionamentos entre entidades. As entidades podem ser atores sociais, páginas web, neurônios do cérebro, dentro outras. Os relacionamentos podem dar-se por meio de trocas materiais (movimentação, proximidade) ou não materiais (informação, sinais elétricos). Em todo os casos, o relacionamento entre entidades pode ser modelados utilizando-se grafos.

Verifica-se assim que a construção do modelo em redes pode gerar frutos futuros em relação à análise do domínio em questão.

Voltando a gravação das tags, para que os dados possam ser armazenados serão necessárias mais três tabelas que deverão ser acopladas ao modelo físico do Dspace: tags, tags2tags e tags2item.

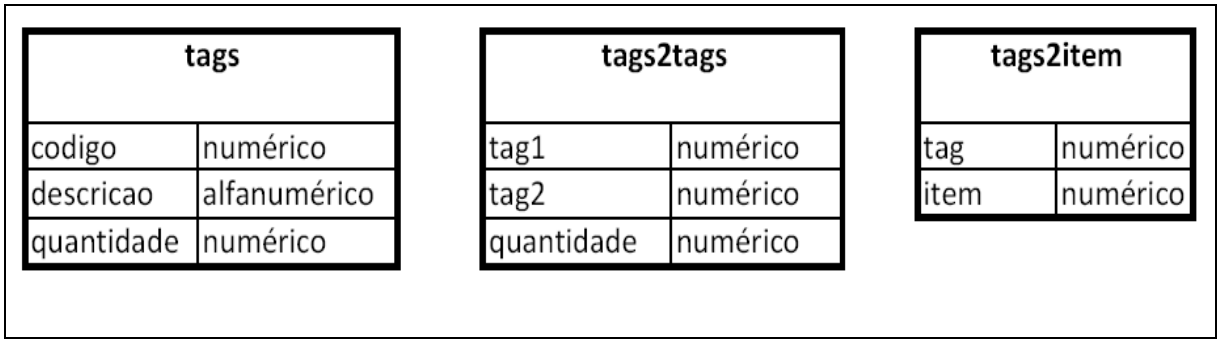


Figura 31 – Tabelas para armazenamento das tags
Fonte: Próprio autor

As tabelas tags e tags2tags (figura 31) serão utilizadas para armazenar os dados referentes às tags descritas no depósito. A tabela tags conta com os campos: código, que indicará um indicador único para cada tag; descrição, que armazenará o texto real da tag; quantidade, que representará a quantidade de vezes que a tag foi utilizada no sistema. A tabela tags2tags indicará nos seus campos tag1 e tag2 os códigos referentes às tags que se relacionam, e o campo quantidade deverá informar a quantidade de vezes que isso acontece.

A tabela tags2item (figura 31) será utilizada para fazer a referência entre os itens (objetos digitais/recurso) armazenados no repositório e as tags que estão diretamente ligadas a eles.

Como exemplo, pode-se utilizar um conjunto de quatro artigos, sendo três deles publicados na revista Datagramazero e outro publicado na revista Brazilian Journal Information Science (BJIS), para demonstrar como ficariam armazenadas as tags na estrutura proposta de tabelas.

Os artigos e suas respectivas palavras-chave, utilizadas como tags neste exemplo, são os seguintes:

- Projeto de ontologia para sistemas de informação empresariais: delineando uma metodologia para desenvolver ontologias na área de telecomunicações, dos autores Beatriz Ainhize Rodriguez Barquín et al., que conta com as seguintes palavras-chave: Ontologia; Sistemas de Informação Empresariais; Web Semântica.
- Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0, dos autores Plácida Leopoldina Ventura Amorim da Costa Santos e Rachel Cristina Vesú Alves, com as seguintes palavras-chave: Informação e Tecnologia; Metadados; Web Semântica; Web 2.0; Web 3.0; Ambientes Informacionais.
- Semelhanças e Diferenças entre Tesauros e Ontologias, dos autores Rodrigo de Sales e Ligia Café, com as seguintes palavras-chave: Tesauro; Ontologia; Linguagem documentária; Representação do conhecimento.
- O nível do conhecimento e os instrumentos de representação: tesauros e ontologias, dos autores Alexandra Moreira, Lídia Alvarenga e Alcione de Paiva Oliveira, com as seguintes palavras-chave: Ontologia; Tesauros; Epistemologia; Representação do Conhecimento.

O conceito de publicação em que a Folksonomia Assistida atua não tem como característica apenas digitar as tags de documentos já publicados, mas evoluir com o processo de caracterização e inserção de tags. Porém, no caso deste exemplo, utiliza-se material já publicado, mostrando, através das figuras 32 e 33, como ficariam registradas no banco de dados essas informações, de forma que possa dar entendimento à construção da estrutura de tabelas sugerida.

tags		
codigo	descricao	quantidade
1	Ontologia	3
2	Sistemas de Informação Empresariais	1
3	Web Semântica	2
4	Informação e Tecnologia	1
5	Metadados	1
6	Web 2.0	1
7	Web 3.0	1
8	Ambientes Informacionais	1
9	Tesauro	2
10	Linguagem Documentária	1
11	Representação do Conhecimento	2
12	Epistemologia	1

Figura 32 – Tabela tags populada

Fonte: Próprio autor

Na figura 32, pode-se verificar que todas as tags foram registradas no banco de dados, sendo que algumas, como o caso de “ontologia”, “web semântica”, “tesauro” e “representação do conhecimento”, são representadas mais de uma vez.

tags2tags		
tag1	tag2	quantidade
1	2	1
1	3	1
2	3	1
3	4	1
3	5	1
3	6	1
3	7	1
3	8	1
4	5	1
4	6	1
4	7	1
4	8	1
5	6	1
5	7	1
5	8	1
6	7	1
6	8	1
7	8	1
1	9	2
1	10	1
1	11	2
9	10	1
9	11	2
10	11	1
1	12	1
9	12	1
11	12	1

tags2item	
tag	item
1	1
2	1
3	1
3	2
4	2
5	2
6	2
7	2
8	2
1	3
9	3
10	3
11	3
1	4
9	4
11	4
12	4

Figura 33 – Tabelas tags2tags e tags2item populadas

Fonte: Próprio autor

Através da figura 33, é possível registrar o armazenamento das relações, sendo que a tabela tags2tags leva a identificar que as tags codificadas como 1 e 9, que representam respectivamente “ontologia” e “tesauro”, estão relacionadas mais de uma vez, assim como 1 e 11, que são “ontologia” e “representação do conhecimento”, também relacionadas mais de uma vez. Essas relações citadas que contemplam mais de uma unidade de relacionamento acontecem, porque as mesmas palavras-chave são utilizadas em mais de um documento.

A tabela tags2item (figura 33) representa a ligação que existe entre as tags e os documentos inseridos, lembrando que ela deve estar relacionada com a tabela item apresentada no modelo físico do Dspace.

6.3.3 Iteratividade, a retroalimentação da informação

A implementação do modelo de Folksonomia Assistida será a base para a consolidação da Representação Iterativa, que deverá ser retroalimentada, sempre baseada no contexto de uma estrutura de representação do conhecimento, através de uma ontologia, taxonomia ou de um tesauro, que consiste em definir os limites de um domínio do conhecimento.

É possível visualizar a Representação Iterativa de forma conceitual. Dada uma visão geral, o modelo é iniciado no usuário, através da extração de informações de um documento, e amparados por estruturas de representação do conhecimento, além de informações já inseridas no sistema por outros usuários, fazem a descrição do objeto digital para efetivar um depósito em um repositório digital científico. As informações cadastradas são utilizadas para amparar o depósito de outros usuários, além de possibilitar a um usuário administrador que, sob observação do conjunto de informações depositadas, faça alterações na estrutura de representação do conhecimento utilizada.

Essa visão geral é detalhada na figura 34 que apresenta os passos para que realmente aconteça o uso completo da Representação Iterativa.

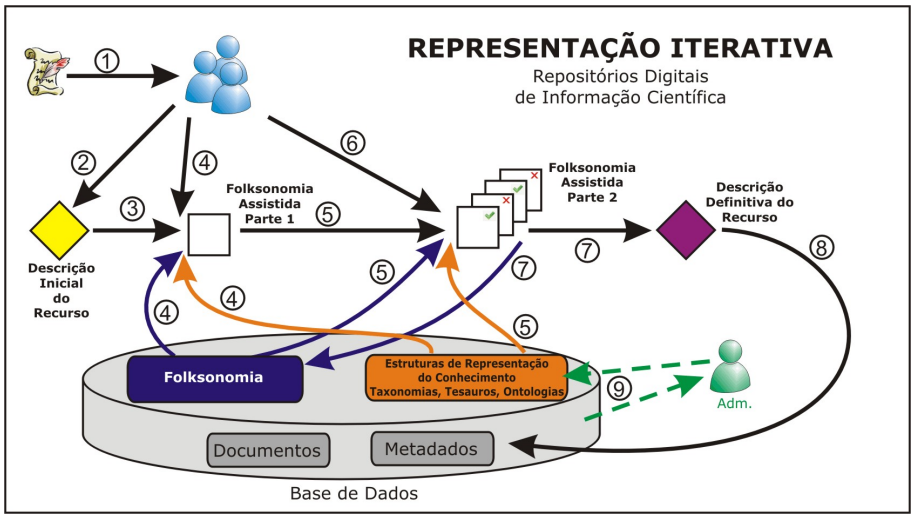


Figura 34 – Representação Iterativa – Visão Detalhada

Fonte: Próprio autor

A construção do modelo nomeado Representação Iterativa, sugerido nesta tese, apresentado de forma detalhada na figura 34, deverá ser construído conforme os seguintes passos:

1. Os usuários fazem uma leitura e verificação do documento a ser depositado e extraem os metadados necessários que descrevam o máximo possível o objeto, para que seja realizado o depósito.
2. O usuário através de formulário disponível no ambiente inicia o processo de descrição do recurso. Esse passo é chamado de descriçao inicial do recurso porque é neste momento em que o usuário deverá inserir todos os metadados relativos ao objeto, com exceção da tag assunto.
3. Com as informações dos metadados já alimentadas, o sistema encaminha o usuário para fazer a descrição da tag assunto, que é a informação que representará de forma mais significativa o recurso dentro da Representação Iterativa.

4. Esse passo representa o início da Folksonomia Assistida. Nesse momento, o sistema deverá colaborar na descrição da tag, utilizando uma estrutura de sugestão, semelhante ao da pesquisa do google, sendo que as informações sugeridas serão os próprios termos já inseridos anteriormente por usuários (Folksonomia – representação livre), além dos termos que fazem parte da estrutura de representação do conhecimemento (taxonomias, ontologias ou tesauros) que estará associada ao repositório.
5. Nesse passo acontece o segundo momento da Folksonomia Assistida. Após a descrição da tag assunto, o ambiente reconhece essas informações e busca relacionamentos e associações dentro do instrumento de estrutura de representação do conhecimento utilizado, agrega termos, e em seguida faz o mesmo dentro do conjunto de tags já definidas por outros usuários (representação livre), busca associações e, na sequência, também relaciona termos. Esse conjunto de termos que foram selecionados são devolvidos para o usuário.
6. O usuário volta a atuar novamente assim que recebe o cojnunto de termos do ambiente. Neste momento ele deve completar o processo de Folksonomia Assistida escolhendo de forma definitiva os termos que serão utilizados na tag assunto. Essa decisão implica em estabelecer relacionamentos entre as tags, e portanto, criar a relação semântica de termos que irá caracterizar a recuperação semântica posterior. Portanto, esse momento é muito importante para a consolidação da Representação Iterativa, porque estabelece os termos e relacionamentos que caracterizam o recurso.
7. Esse passo apenas apresenta a confirmação da descrição completa do recurso, visto que o usuário já descreveu inicialmente os metadados e em seguida, com auxílio da

Folksonomia Assistida, escolheu os termos que compõe a tag assunto. É nesse momento que a Folksonomia (representação livre) será alimentada efetivamente com o novo conjunto de termos e relacionamentos que o usuário efetivou e dessa forma reorganizada, atualizando o peso dos termos e relacioanamentos de acordo com os novos elementos que foram inseridos. Cada vez que esse passo é efetivado em um novo depósito acontece um enriquecimento e fortalecimento do conjunto de termos e relações existentes, e as informações que foram inseridas passam a ficar disponíveis para serem utilizadas por novos usuários em novos depósitos.

8. Nesse passo o conjunto completo de metadados assim como os objetos digitais são armazenados na base de dados.
9. A cada período de tempo, o processo deverá ser avaliado por um administrador de sistema que poderá também retroalimentar o a estrutura de representação do conhecimento das áreas de especialidades, dando uma nova visão a respeito dos limites estabelecidos ao domínio do conhecimento. Esse processo cria uma nova perspectiva na Ciência da Informação, que é a avaliação e reconstrução da estrutura de representação do conhecimento, baseado na construção da informação, por usuários de um ambiente digital.

É importante ressaltar que o administrador deve ser um profissional ou equipe multidisciplinar responsável pela catalogação do ambiente informacional e pela manutenção das estruturas de representação do conhecimento (bibliotecário, arquivísta e/ou cientista da informação).

O processo de iteratividade é estabelecido de forma que fica a cargo de um usuário administrador a retroalimentação da estrutura de representação do conhecimento, e, como função sistemática e automática dos usuários, as retroalimentações da Folksonomia.

O processo de iteratividade resulta na reconstrução do conhecimento, de forma coletiva e moderada, permitindo o enriquecimento e amadurecimento da estrutura de representação do conhecimento para o domínio em que o repositório digital científico está inserido.

A arquitetura proposta neste trabalho parte do princípio da iteratividade, que é o processo em que ocorre a realimentação constante do sistema em busca da melhor qualidade do conjunto de informações.

O princípio da iteratividade está dentro do contexto de desenvolvimento de software, do qual foi realizada uma adaptação para a construção deste modelo estrutural para repositórios digitais científicos. É importante ressaltar que o estudo de processos e metodologias para melhorar o desenvolvimento de software é constante dentro da área de Ciência da Computação.

Jacobson et al. (1999) afirma:

> O processo de desenvolvimento de software é um conjunto de atividades e resultados associados que tem por objetivo produzir software eficiente, de alta qualidade, com baixa taxa de erros e que atenda às necessidades e expectativas do usuário de forma geral.

O conceito de desenvolvimento iterativo é bastante utilizado na Engenharia de Software, disciplina da Ciência da Computação, e faz parte de alguns processos de desenvolvimento de software já sedimentados e muito utilizados como RUP (Rational Unified Process), Programação Extrema (XP) e Scrum.

Segundo Larman (p. 47, 2007),

> O ciclo de vida iterativo é baseado em refinamentos e incrementos sucessivos de um sistema por meio de múltiplas iterações, com realimentação (feedback) e adaptação cíclicas como principais propulsores para convergir para um sistema adequado. O sistema cresce incrementalmente ao longo do tempo, iteração por iteração, razão pela qual esta abordagem também é conhecida como desenvolvimento iterativo e incremental. Como a realimentação e adaptação fazem as especificações e o projeto evoluir, esse sistema é conhecido como desenvolvimento iterativo e evolutivo.

Larman afirma que o processo iterativo é também evolutivo, assim como acontece no modelo proposto nesta tese, que propõe a evolução das representações de informação.

Essa evolução pode ocorrer através da Folksonomia Assistida, ou seja, do processo repetitivo de inserção de conteúdos para tags, assim como da evolução e adaptação da estrutura de representaç ao do conhecimento utilizada, por intermédio de um administrador.

O processo de desenvolvimento iterativo, do qual a Representação Iterativa é baseada, deve obedecer limites temporais.

Larman (2007, p. 50) alerta sobre os limites temporais:

> A maioria dos métodos iterativos recomenda que a duração de uma iteração seja entre duas e seis semanas. Usar pequenos passos, obter realimentação rápida e fazer adaptações são idéias centrais no desenvolvimento iterativo; iterações longas subvertem a motivação central para o desenvolvimento iterativo e aumenta o risco do projeto.

A Representação Iterativa tem um contexto diferente, porque não trata de desenvolvimento de software, mas sim da construção do corpus de informação de um domínio, através de uma inteligência coletiva, porém o princípio da temporalidade também pode ser abordado e utilizado.

O processo de construção da inteligência coletiva pela Folksonomia Assistida não deve ser temporal, ele deve ser feito dinamicamente, sugerindo que o usuário possa ter acesso ao conjunto de informações a qualquer momento, ou seja, assim que uma tag é inserida no sistema, ela passa a ficar disponível para ser utilizada como sugestão a outros usuários. O acesso do administrador do sistema, para fazer ajustes ao modelo que está sendo construído, pode ter sim uma temporalidade definida, corroborando com a ideia de iteratividade. Esta pesquisa não define um intervalo exato de temporalidade de intervenção do administrador do sistema, porém cada ambiente deve estabelecer seu próprio intervalo de temporalidade de acordo com o a quantidade de acessos e o volume de informações dentro do repositório.

Cada iteração gera um novo conjunto de informações, relacionamentos e também uma forma diferente de conhecimento.

Dentro do contexto evolutivo do modelo, é possível que a interatividade entre os usuários e o sistema gere uma camada de informações cada vez mais rica, principalmente porque permite *feedback* ao usuário, assim como a possibilidade da informação já armazenada serve como base para que a próxima seja inserida.

Dessa forma, a Representação Iterativa oferece aos repositórios um novo formato de organização da informação, de modo que passe a existir uma relação entre os trabalhos autoarquivados, não apenas pela simples sintaxe das palavras-chave e nem tampouco pela comunidade e coleção de que fazem parte.

A estrutura funcional deste modelo parte do princípio da agregação de valores ao repositório, de forma que ocorra uma contextualização do material digital inserido, criando relações que possam sustentar uma recuperação semântica de informações.

O processo de recuperação, baseado nesse novo modelo de representação da informação, será abordado no próximo capítulo.

7 RECUPERAÇÃO DA INFORMAÇÃO NO MODELO DE REPRESENTAÇÃO ITERATIVA

O capítulo anterior descreveu o procedimento e modelo criado para construir uma estrutura de repositórios que contemple tecnologias de Web 2.0 e Web 3.0, denominado Representação Iterativa. A construção e a aplicação deste modelo alteram a estrutura dos repositórios digitais e permitem que seja revisto o conceito de recuperação utilizado nesse tipo de ambiente.

Baseado no modelo Representação Iterativa, este capítulo tem como contexto a apresentação de novos métodos de recuperação para repositórios digitais, baseado na utilização de funcionalidades da Web 2.0 e da Web 3.0.

Todo processo anterior foi construído com o objetivo de permitir a recuperação semântica, pois, para que exista uma recuperação baseada em conteúdo, é necessário que exista uma estrutura de armazenamento e descrição da informação, conforme o modelo proposto.

A recuperação semântica pauta do princípio de que não ocorrerá recuperação da informação apenas por comparação sintática de caracteres através do termo inserido pelo usuário no momento da busca, e tão somente por objetos textuais.

Santarem Segundo (2004, p. 16) afirma:

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

A Representação Iterativa vem no sentido de colaborar justamente com a recuperação da informação, independente do formato em que ela estiver.

Segundo Buckland (2006, p.6),

> A técnica de pesquisa por seqüências de caracteres de texto funciona muito bem, mas nem sempre e não perfeitamente,

> porque recursos de texto não são inteiramente homogêneos. Algumas palavras possuem vários significados (polissemia, por exemplo, mouse); às vezes palavras diferentes utilizam a mesma seqüência de caracteres, mas com outros significados (homógrafos, por exemplo, pane significa painel de vidro em inglês, mas não em português); e palavras diferentes podem ser utilizadas com o mesmo significado (sinônimos, por exemplo, câncer e neoplasma).

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

Segundo o documento Buscando termos perto de outros (2003), publicado no site do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico),

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

Alguns mecanismos de busca na Web disponibilizam o recurso de proximidade, porém não é comum o uso desse operador.

7.1 Critérios para recuperação da informação na Representação Iterativa

A proposta de recuperação apresentada nesta pesquisa segue o modelo inicialmente proposto pela própria ferramenta Dspace, oferecendo ao usuário a recuperação através da digitação de um termo e solicitando a pesquisa através de um dos campos escolhido pelo usuário. Além desse formato, também poderá ser oferecida a nuvem de tags, que deverá ser

formulada com base no conjunto de tags inseridas pelos usuários do ambiente digital.

Além dos métodos já citados para a recuperação da informação, também deve ser proposto como ferramenta de busca uma rede de relacionamentos criada através das tags. Essa rede de relacionamentos necessita de uma implementação gráfica, mas garante ao usuário uma navegação entre as tags que estão relacionadas.

O sistema de recuperação da informação para Representação Iterativa deverá seguir os seguintes passos:

- O processo de digitação do termo a ser procurado deverá ser agregado ao oferecimento sugestivo de termos que compõem as tags já cadastradas no sistema, se esse for inserido de forma digitada pelo usuário. A outra forma é através da nuvem de tags ou então da rede de tags.
- O sistema deverá receber essa informação (termo) e buscar de forma sintática a relação entre o termo digitado e o conjunto de informações que compõe a base de dados de tags, assim como proceder à mesma busca, de forma sintática, na estrutura de representação do conhecimento das áreas de especialidades, que deverá estar associado ao repositório.
- Ao encontrar uma referência sintática, deverá então, baseado na estrutura de informação construída, buscar as relações semânticas que existem no modelo para o termo digitado pelo usuário e construir um novo conjunto de informações com novos termos, porém relacionados semanticamente ao primeiro, e novamente submeter a pesquisa sintática ao conjunto de objetos cadastrados no repositório digital. Apesar de a busca ser estritamente por palavras-chave, pode ser estendida para procurar os termos no título e subtítulo dos documentos depositados.

- Essa nova pesquisa, com a agregação de termos que foram relacionados sem utilizar comparação sintática, deverá oferecer resultados que tenham como base o relacionamento vertical e horizontal dos termos, finalizando o processo de recuperação da informação.

A apresentação do resultado ao usuário deverá ser feito de forma que os termos que geraram o resultado apareçam inicialmente e, em seguida, todos os links gerados por aquele termo, e assim, sucessivamente, até que se esgotem os termos agregados a esta busca, conforme se pode observar na figura 35.

A partir do momento em que o usuário selecionar o resultado apresentado, seguindo para a visualização completa do item, a sugestão é que o item venha apresentado no formato padrão que o Dspace e outras ferramentas já oferecem, com a informação do metadado completo ou parcial e também com a opção de download dos arquivos que compõe o item.

Neste caso indica-se que a apresentação do resultado seja contemplada com a descrição da informação também no formado de microformatos, utilizando-se do microformato DC, de modo que a informação do item possa ter uma estrutura que permita ser identificada e utilizada de forma automática pelo browser que o usuário estiver utilizando.

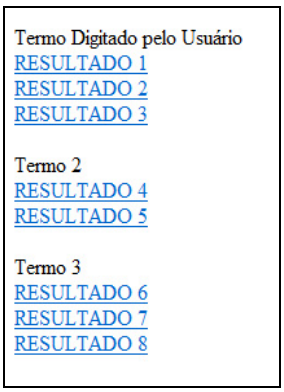


Figura 35 – Exemplo de página de resultados.
Fonte: Próprio autor

A recuperação da informação neste contexto deve seguir uma metodologia que procure garantir que os resultados sejam os mais apropriados para o usuário. Dessa forma, é necessário que as resultados sigam critérios de valoração baseados nas informações internas.

Os critérios estabelecidos para promover a apresentação dos resultados são:

- Formar, primeiramente, um grupo de termos que foram estabelecidos como apropriados após recuperação dentro do conjunto de tags e da estrutura de representação do conhecimento, sendo que estes deverão estar em ordem de preferência. Ou seja, será construída uma lista ordenada de termos.
- A montagem da lista deverá ser encabeçada pelo termo digitado pelo usuário; na sequência, pelos termos que tiverem relacionamento de um nível, vertical ou horizontal, dentro da estrutura de representação do conhecimento das áreas de especialidades; posteriormente, pelos termos que apresentem maior densidade de relacionamento através do cruzamento de tags com o termo digitado pelo usuário. No caso de utilizar densidade do relacionamento entre tags, caso haja valoração igual, o "desempate" deverá vir através das tags que foram mais citadas no sistema.
- Após a confirmação da lista ordenada, o processo de apresentação da informação terá como prioridade mostrar os documentos que contenham em seu conjunto de tags o termo escolhido pelo usuário. Caso haja mais de um registro que contenha o termo, então deverá ser verificado, na lista ordenada de tags, se os documentos têm alguma outra tag desta lista, e, se houver, deverá ser dada prioridade maior ao documento que contiver as tags que aparecem primeiro na lista ordenada de tags.

- A sequência de apresentação de resultado deverá ser procedida de forma que sejam verificadas, nos documentos, as tags que figuram nas posições superiores na lista ordenada. Neste caso, se houver documentos que, utilizando os critérios estabelecidos até então, continuem "empatados", deverá ser apresentado primeiro o documento que apresentar o maior número de relacionamento de suas tags com outras de forma geral no sistema, indicando que este documento está "mais relacionado" com o domínio do conhecimento do que o outro.

O nível de relacionamento entre o termo digitado e os termos recuperados para proceder à apresentação do resultado, apesar de sugerido como "1" nesta pesquisa, poderá ser parametrizado pelo administrador do sistema, ou ainda, definido pelo usuário no momento da pesquisa. Se essa definição ficar a cargo do usuário, em breve ele verificará que quanto menor for o valor estabelecido para relacionamento, mais fechada e coesa ficará sua pesquisa, e, ao contrário, maior será a quantidade de resultados apresentados.

Guy e Tonkin (2006, p.3) dizem que,

> Há uma série de ferramentas disponíveis que oferecem uma variedade de métodos de visualização diferentes para sistemas que usam Folksonomia, principalmente o Del.icio.us, incluindo tag.alicio.us[33], extisp.icio.us[34] e jocoso[35].

Portanto, é importante que possamos oferecer mais modelos de recuperação da informação aos usuários dos repositórios que contemplam Representação Iterativa. Neste caso, modelos gráficos, como nuvem e rede de tags são ainda mais intuitivos, facilitando o processo de recuperação pretendido pelo usuário.

---

[33] http://planetozh.com/blog/2004/10/05/tagalicious-a-way-to-integrate-delicious/
[34] http://kevan.org/extispicious
[35] http://www.siderean.com/delicious/facetious.jsp

7.2 Nuvem de tags

A nuvem de tags, ou tag clouds, como tem sido chamado esse recurso, é uma implementação funcional que permite aos usuários de um ambiente digital verificar visualmente o conjunto de tags que mais estão sendo citadas dentro de um ambiente.

No modelo de Representação Iterativa, que prima pela utilização de Folksonomia, a implementação de uma nuvem de tags é fundamental na apresentação visual do repositório que implementa o modelo.

A apresentação da nuvem de tags, além de promover a visualização das tags mais citadas no repositório, ainda garante ao usuário, através de um simples clique, a recuperação de documentos que estão ligados ao termo que foi clicado.

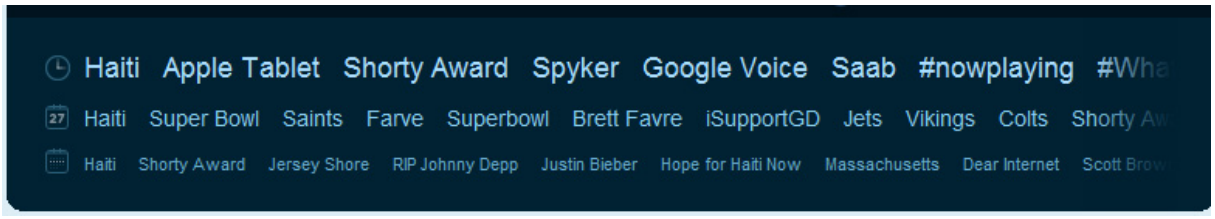


Figura 36 – Nuvem de tags do microblog Twitter
Fonte: http://www.twitter.com

A nuvem de tags, para os repositórios baseados no modelo sugerido neste trabalho, não deve utilizar temporalidade para estabelecer a representação das tags mais utilizadas, portanto, a nuvem de tags terá como base todo o período de utilização do repositório.

É normal verificar ferramentas que, além das tags mais populares, de modo geral, também apresentem as tags mais populares em determinados períodos. A figura 36 apresenta a nuvem de tags do microblog Twitter, separada em três linhas horizontais: a primeira apresenta as tags mais populares do momento atual; a segunda linha, as tags mais populares do dia; e a ultima linha, as tags mais populares da última semana.

O processo tradicional de construção da nuvem de tags estabelece que em uma determinada área do portal ou site, neste caso, da página principal do repositório, deve ser apresentada a nuvem de tags. O processo de apresentação da informação deverá oferecer um grupo de palavras, em uma quantidade inicialmente estabelecida, que, no caso dos repositórios, pode ser definida por volta de 25, de forma que estas palavras tenham tamanhos e tipologia diferentes, de acordo com o nível de destaque e popularidade que ela represente para o repositório.

O estabelecimento de uma quantidade de palavras para compor a nuvem de tags está baseada no espaço reservado para a nuvem de tags dentro do repositório, na página principal, de forma que as tags possam ter tamanhos satisfatórios para telas que usem resolução 1024x768.

A construção da nuvem de tags que representa de forma adequada a Representação Iterativa deverá seguir os seguintes critérios:

- As tags terão a mesma cor e tipo de letra, apresentando apenas diferença no tamanho da letra;
- As tags serão divididas em cinco níveis de apresentação, ou seja, cinco tamanhos diferentes de letras entre as tags apresentadas.

O primeiro passo será recuperar no banco de dados, na tabela “tags”, definida na Representação Iterativa, os 25 termos mais citados. A informação relativa à quantidade de ocorrências do termo não está relacionada à quantidade de relacionamentos estabelecidos pelo termo, mas sim pelo número de documentos que citam o termo como tag. O campo “quantidade” da tabela “tags” tem a informação da quantidade de vezes em que o termo foi citado.

Após recuperar os 25 termos mais citados, deverá ser calculado um número que servirá como guia para estabelecimento de cada um dos 5 níveis em que as tags estarão divididas. Cada um desses níveis deverá representar um tamanho de fonte diferente.

Para calcular o número guia, que representará o valor do intervalo de cada nível, o procedimento adotado é utilizar a quantidade de vezes do termo mais citado e subtrair a quantidade do termo menos citado, e, na sequência, dividir o resultado pela quantidade de níveis que a nuvem de tags terá, nesse caso o valor cinco.

Exemplo: caso o termo mais citado seja "ontologia", com 70 ocorrências, e os termos menos citados (plural porque muitas vezes há mais de um termo com a quantidade mínima) sejam com 2 ocorrências, então se terá como número guia o valor 13,6, resultado da subtração de 2 ocorrências dos termos menos citados das 70 ocorrências do termo "ontologia", dividido pelo valor 5, que representa a quantidade de níveis estabelecidos para o sistema.

Após a definição do valor guia, em 13,6, deverá ser estabelecido o limite dos níveis em que as tags estarão dispostas.

Portanto o modelo matemático para se estabelecer o intervalo entre os nível, chamado de número guia, é a seguinte: **g = (T – t) / ns**, onde:

- **g**: significa o número guia, ou seja, o intervalo que deverá ocorrer entre os níveis.
- **T**: é a quantidade de ocorrências do termo mais citado no conjunto de termos selecionados.
- **t**: é a quantidade de ocorrência dos termos menos citados no conjunto de termos selecionados.
- **ns**: é a quantidade de níveis que se deseja utilizar na nuvem de tags.

Para definir o intervalo dos níveis outro modelo matemático deverá ser utilizado, porém o primeiro nível terá seu valor inicial estabelecido de acordo com a quantidade de ocorrência das tags menos citadas, dessa maneira estabeleceremos que o modelo matemático que compreende os níveis são: **lin <= n < lsn**, sendo que o primeiro nível **lin=t**, e a partir dos

próximos níveis **lin** do próximo nível será igual ao **lsn** do nível anterior, onde:

- **lin**: limite inferior do nível;
- **n**: nível a que estão sendo estabelecidos os limites;
- **lsn**: limite superior do nível;

O modelo matemático para calcular o lsn é: **lsn = (t + g * n),** onde:

- **lsn**: limite superior do nível;
- **t**: quantidade de ocorrência dos termos menos citados no conjunto de termos selecionados;
- **n**: é o nível a que se está estabelecendo o calculo;
- **g**: número guia calculado no primeiro modelo matemático apresentado.

Dessa forma, continuando o exemplo:

- o nível 1 deverá ter como limite inferior (lin) o valor 2, e como limite superior (lsn) o valor = 15,6 => (2 + 13,6 * 1);
- o nível 2 deverá ter como limite inferior (lin) o valor 15,6, e como limite superior (lsn) o valor = 29,2 => (2 + 13,6 * 2);
- o último nível, nesse exemplo, deverá ter como limite inferior (lin) o valor 56,4 , e como limite superior (lsn) o valor 70 => (2 + 13,6 * 5).

Portanto, o primeiro nível será caracterizado pelas tags que apareceram menos do que 15,6 vezes; o segundo nível será estabelecido entre as tags que foram citadas entre 15,7 até 29,2 vezes, e assim por diante até formar o último nível, com valor teto de 70 ocorrências da tag, que deve ser igual ao valor da tag com maior frequência.

Após delimitar os 5 níveis de apresentação das tags, deverá ser escolhido um tamanho de fonte que represente cada um dos 5 níveis e fazer uma leitura sequencial, alfabética ou aleatória das 25 tags mais citadas.

Conforme o nível que ela estiver inserida, deverá ser apresentada com um tamanho de fonte correspondente ao nível.

Esse formato de criação de nuvem de tags é um dos mais utilizados nos portais de Internet, e há diversos scripts disponíveis com sugestão de criação de nuvem de tags, nos mais diversos fóruns de discussão a respeito de desenvolvimento de sistemas para Internet.

Acredita-se que a utilização de 25 tags e dos 5 níveis deverá corresponder à estrutura de um repositório, porém no início esses valores deverão ser reduzidos, e, posteriormente, de acordo com a frequência de utilização do repositório, poderá também ser expandido.

Após a apresentação da nuvem de tags, o recurso ficará disponível para cliques dos usuários. Assim, toda vez que ocorrer o clique do usuário, o sistema deverá iniciar o processo de busca, conforme apresentado no início desse capítulo, prevendo uma recuperação semântica de informação para o atendimento das necessidades do usuário, e com apresentação dos resultados (figura 35).

7.3 Rede de tags

A estrutura da Representação Iterativa permite criar um novo sistema de recuperação da informação dentro dos repositórios. O novo modelo não deve substituir o anterior, mas sim agregar mais um tipo de pesquisa e interação do usuário com o ambiente.

O formato de rede tem sido muito abordado no conceito de colaboração científica, principalmente nos relacionamentos entre coautorias e cocitações, porém a mesma ideia utilizada neste conceito se aplica às redes de tags, que podem agregar a informação a respeito dos autores e criar o conceito de autores que tenham o mesmo perfil de depósito dentro de um repositório.

Segundo Wasserman e Faust (1994, p.9),

> o termo 'rede social' se refere ao conjunto de atores e suas ligações entre eles. Assim, a análise de rede tem por objetivo modelar as conexões entre os atores, a fim de retratar, descrever e representar a estrutura de um grupo, quer seja composto por países, instituições ou pessoas.

O modelo em formato de rede aproxima termos que estão relacionados criando uma estrutura de informação que tem apresentação visual agradável e de entendimento intuitivo.

Tannuri e Gracio (2008, p. 39) afirmam:

> As análises métricas oferecem subsídios e instrumentação para o estudo das redes sociais na medida em que, a partir de tratamentos quantitativos, torna possível a avaliação de alguns aspectos dessas relações, através de gráficos, densidades, proximidades, similaridades, vetores, intensidades, centralidades e homogeneidades. Assim, a ligação entre dois pontos pode significar não só a existência da colaboração científica entre autores e instituições científicas, mas também a intensidade dessa colaboração na forma de co-autorias.

Portanto, se a ideia de coautoria das redes colaborativas torna possível uma grande quantidade de estudos, as redes construídas através da estrutura da Representação Iterativa poderão gerar um conjunto grande de informações a respeito do conteúdo dos objetos depositados nos repositórios digitais informacionais.

O estudo a respeito da análise das redes que a Representação Iterativa proporciona não faz parte dos limites desta pesquisa, porém pode ser considerado como trabalho futuro.

Para a criação de uma rede de informações que permita ao usuário navegar pelos termos, os dados registrados nas tabelas "tags" e "tags2tags" deverão formar uma matriz de adjacência que possibilite a construção do grafo, que é a estrutura matemática e computacional escolhida para representar as redes.

O exemplo apresentado no capítulo 6, a respeito dos quatro artigos que geraram um grupo de informações para compor as tabelas físicas do repositório, será aproveitado aqui para compor a rede de tags.

O primeiro passo para a construção da rede de tags é gerar uma matriz de adjacência, que dá sustentação à criação do grafo/rede.

A matriz de adjacência é construída de forma que as linhas e colunas da matriz sejam representadas pelas tags e o cruzamento indica a quantidade de relacionamentos existentes entre as tags.

A matriz de adjacências baseada no exemplo anterior é apresentada na figura 37.

	Ontologia	Sistemas de Informação Empresariais	Web Semântica	Informação e Tecnologia	Metadados	Web 2.0	Web 3.0	Ambientes Informacionais	Tesauro	Linguagem Documentária	Representação do Conhecimento	Epistemologia
Epistemologia	1								1		1	
Representação do Conhecimento	2								2			
Linguagem Documentária	1								1			
Tesauro	2											
Ambientes Informacionais			1	1	1	1	1					
Web 3.0			1	1	1	1						
Web 2.0			1	1	1							
Metadados			1	1								
Informação e Tecnologia			1									
Web Semântica	1	1										
Sistemas de Informação Empresariais	1											
Ontologia												

Figura 37 – Matriz de adjacências e quatro artigos utilizados como exemplo.
Fonte: Próprio autor

Baseado na matriz de adjacências construída é possível construir o grafo de tags.

A apresentação do grafo possibilita algumas variações, e neste trabalho sugere-se que os vértices tenham tamanhos diferentes, de forma proporcional, utilizando a mesma técnica de construção da nuvem de tags, através da construção de um valor guia e definições de níveis. A diferença em relação à nuvem de tags é que, no modelo de redes, todas as tags deverão

fazer parte do grafo, mesmo que apenas parte dela seja apresentada ao usuário.

Portanto pode-se definir também 5 níveis de apresentação dos termos (círculos), sendo que cada nível terá um tamanho diferente, ficando os temos mais populares com o maior diâmetro e os menos populares com menor diâmetro, conforme figura 38.

As arestas que ligam os vértices e que representam a quantidade de ligações existentes entre cada uma das tags também deverão seguir o padrão proposto na nuvem de tags. Então, a quantidade de relacionamentos existentes entre as tags será representada visualmente pela largura da ligação entre as arestas, e a largura das relações deverá ser construída com o emprego de níveis pré-estabelecidos, como é realizado na nuvem de tags, ou seja, quando maior o peso entre dois termos mais larga será a linha que une os termos, e quanto menor o peso mais fina será a linha, conforme pode ser visto na figura 38.

O modelo de Representação Iterativa sugere a mesma quantidade de níveis utilizada na nuvem de tags para a definição dos níveis dos relacionamentos entre os termos, que na verdade representam o peso de um relacionamento entre dois termos.

No plano de visualização da rede pelo usuário, é inviável que seja apresentada toda a rede de tags, portanto a Representação Iterativa sugere a apresentação de termos que estejam a uma distância (d) de dois ou três termos do termo que é apresentado como termo (nó) principal da rede de tags, porém, à medida que o usuário vai navegando na rede, o nó principal passa a ser trocado e então mudam a profundidade e largura, para que novos vértices do grafo passem a fazer parte da visualização. O procedimento de apresentação deverá ser calculado através do procedimento de busca em largura e busca em profundidade.

A distancia (d) entre termos é a quantidade de nós que deve-se passar para se chegar de um termo a outro.

Quando o usuário proceder com dois cliques em um nó da rede, então deverá ser executado o procedimento de recuperação e apresentação dos resultados, conforme procedimento de busca e apresentação de resultado (figura 38).

A definição do nó principal da rede deve se dar através do termo que é mais citado no repositório, iniciando a rede sempre por esse termo.

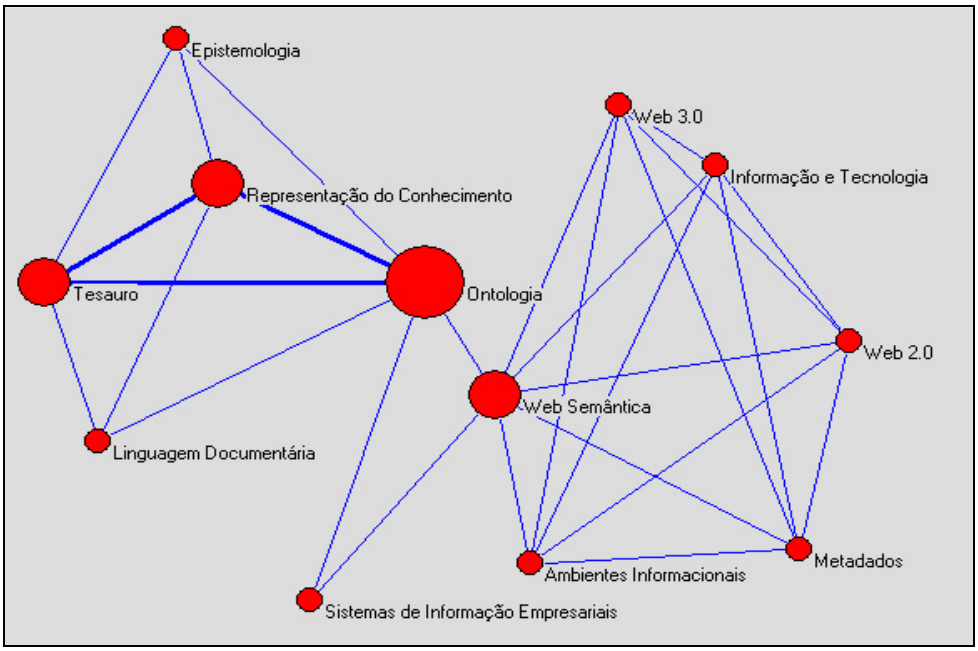


Figura 38 – Rede de Tags de quatro artigos utilizados como exemplo.
Fonte: Próprio autor

O grafo resultante da matriz apresentada na figura 37 pode ser visualizado na figura 38, porém, dado o suporte de apresentação deste trabalho, não é possível realizar o deslocamento, que deverá ser implementado através de técnicas de programação visual no ambiente.

Dessa forma, apresentam-se novos conceitos de recuperação da informação, baseados na Representação Iterativa. Assim, finda-se o trabalho com a completude de um modelo que pode mudar a estrutura funcional dos repositórios digitais, de forma a permitir que estes sejam ambientes mais ricos e aptos a construir a recuperação semântica de informações.

8. CONCLUSÕES

A construção desta pesquisa nasceu da necessidade de melhorar a recuperação da informação em repositórios digitais informacionais. Para que isso fosse possível, foi proposto um modelo novo nomeado Representação Iterativa para repositórios digitais.

O primeiro passo para iniciar a construção do modelo proposto foi verificar que os repositórios instituicionais são ambientes que necessitam de melhorias, tanto do ponto de vista das funcionalidades oferecidas aos usuários, quanto do ponto de vista técnico para a recuperação da informação.

Assim, algumas considerações a respeito da estrutura foram evidenciadas, como a falta de funcionalidades que são implementadas pelos grandes portais, como os recursos que a Web 2.0 oferece. Dentre os recursos da Web 2.0, foi encontrado o RSS, que, de certa forma, tem um nível baixo de complexidade e implementação.

Foi possível verificar também que as ferramentas disponíveis para implementação de repositórios foram construídas sobre um modelo que oferece condições de implantação dos recursos da Web Semântica. Alguns pontos são fundamentais, como a utilização do formato de metadados Dublin Core.

Verificou-se ainda que a Folksonomia é um processo importantíssimo para ser aplicado ao contexto dos repositórios, visto que permite a construção de inteligência coletiva e oferece subsídios para que haja uma busca por termos relacionados, porém se for efetivamente utilizado de forma totalmente livre pode gerar termos sem relacionamentos futuros, ou ainda inexatos e inconsistentes dentro da Representação Iterativa.

Conclui-se que a necessidade de um novo conceito de Folksonomia, a Folksonomia Assistida, proposta neste trabalho, vem ao encontro à necessidade de auxiliar o usuário na descrição da tag assunto do recurso a ser depositado, em relação ao domínio do conhecimento do qual o

repositório faz parte. A Folksonomia Assistida é um processo que pode definitivamente elevar o nível de qualidade de descrição do recurso, de forma que relacionam os itens depositados a termos que estão no pensamento e conhecimento dos usuários do sistema.

A Folksonomia Assistida é um processo de auxílio na descrição do recurso e fundamental na elevação da qualidade da descrição do recurso, mantendo a criatividade do usuário na inserção da tag, mas também oferecendo a ele elementos que possam relacionar seu recurso a outros já depositados anteriormente ou/e ainda a uma estrutura de representação do conhecimento.

A construção do novo modelo permitiu agregar funcionalidades importantes ao repositório, possibilitando a recuperação da informação.

O modelo de Representação Iterativa, principal proposta deste trabalho, é de fundamental importância no papel de qualificar e melhorar a estrutura de representação do conhecimento das áreas de especialidades, visto que, do ponto de vista da evolução, uma estrutura de representação do conhecimento pode encontrar subsídios na utilização das tags propostas no sistema para melhor se adequar ao domínio e aos pesquisadores que utilizam o repositório.

O modelo de Representação Iterativa estabelece peso entre os termos inseridos na tag assunto, fortalecendo a relação entre termos que tem relação semântica e estabelecendo uma ligação entre estruturas de representação livre (Folksonomia) e estruturas de representação do conhecimento (Ontologias, Tesauros e Taxonomias), criando assim um ambiente definitivamente semântico de construção colaborativa.

A participação do usuário no modelo Representação Iterativa é fundamental, visto o perfil do usuário pode condicionar ao bom funcionamento da Representação Iterativa.

A estrutura de banco de dados elaborada garante a construção de ferramentas que tendem a melhorar muito o processo de recuperação semântica da informação, como a nuvem e a rede de tags.

Para que ocorra a recuperação da informação dentro de um contexto semântico, deve haver uma estrutura de armazenamento que sustente isso.

Verificou-se também que desenvolver modelos gráficos para amparar a recuperação da informação pode facilitar e auxiliar os usuários no processo de recuperação da informação em ambientes que utilizam-se de Folksonomia, como a Representação Iterativa.

Dentro do contexto da Representação Iterativa, verifica-se que os modelos vetorial e genético de recuperação da informação, podem contribuir muito no contexto global de recuperação da informação, visto que a relação de peso entre as ligações, que existe no modelo vetorial, e a retro-alimentação da informação com participação do usuário, utilizando-se de um refinamento de acordo com o ambiente, contribuem para o contexto de recuperação semântica da informação.

Conclui-se também que o modelo não é restrito a repositórios digitais, apesar de ter sido o foco da pesquisa. A Representação Iterativa e a Folksonomia Assistida podem ser aplicadas em outros tipos de ambientes digitais que ofereçam ao usuário a possibilidade de descrever suas próprias tags e trabalhem com uma estrutura de representação do conhecimento das áreas de especialidades.

8.1 Projetos Futuros

O modelo Representação Iterativa abre as portas para que novas pesquisas possam ser realizadas, a principal delas é a implementação técnica do modelo.

Oferecer uma estrutura que possa armazenar o peso das ligações entre a Folksonomia (representação livre) e as estruturas de representação do conhecimento também pode ser abordado, de forma que aumente a

relação semântica entre essas duas estruturas de informação e conhecimento.

Analisar os resultados e o conjunto de informações armazenadas dentro desse novo contexto de repositório também pode agregar mais valor a esta pesquisa, visto que permite avaliar se colabora efetivamente com a iteratividade de atualização de uma estrutura de representação do conhecimento.

Aplicar o conceito de Representação Iterativa em outros tipos de ambientes, construir redes de colaboração utilizando autores, baseadas nas tags que eles utilizam, também poderão contemplar a Ciência da Informação, com a dimensão em que os pesquisadores atuam e, de certa forma, trabalham dentro de um mesmo domínio.

Faz-se necessário construir um novo modelo que interfira na Representação Iterativa de modo a analisar os resultados apresentados e os resultados utilizados pelo usuário, de forma que isso possa alterar as relações de termos criados no depósito dos objetos digitais.

REFERÊNCIAS

ALMEIDA, R. L. de. Da disseminação seletiva à web syndication: uma proposta para a comunicação científica. In: Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB, 8., 2007, Salvador. **Anais eletrônicos**... Salvador: ANCIB, 2007. Disponível em: <http://www.enancib.ppgci.ufba.br/artigos/GT7--157.pdf>. Acesso em: abr. 2009.

ALVES, R. 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.

ANSI Z39-19-2005. **Guidelines for the construction, format, and management of monolingual controlled vocabularies**. Bethesda: NISO Press, 2005.

AQUINO, M. C. Hipertexto 2.0, folksonomia e memória coletiva: um estudo das tags na organização da web. **E-Compós**, Brasília, v. 9, 2007. Disponível em:< http://www.compos.org.br/seer/index.php/e-compos/article/view/165/166>. Acesso em: 3 nov. 2009.

ARAUJO, M. de. **Educação a distância e a Web Semântica:** modelagem ontológica de materiais e objetos de aprendizagem para a plataforma COL**.** 2003. 173f. Tese (Doutorado) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2003. Disponível em:< www.teses.usp.br> Acesso em: maio 2008.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM; Harlow: Addison-Wesley, 1999.

BARQUÍN, B. A. R. et al. Projeto de ontologia para sistemas de informação empresariais: delineando uma metodologia para desenvolver ontologias na área de telecomunicações. **Brazilian Journal of Information Science**, Marília, v.2, n. 2, p. 17-34, jul./dez. 2008.

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

BECHHOFER, S. et al. **OWL Web Ontology Language reference.** 2004. Disponível em: < http://www.w3.org/TR/owl-ref/>. Acesso em: 22 maio 2009.

BEKAERT, J.; VAN DE SOMPEL, H. **Augmenting interoperability across scholarly repositories**. Report, 2006. Disponível em: <http://msc.mellon.org/Meetings/Interop/FinalReport >. Acesso em: 14 fev. 2009.

BENTLET, P. J. **Biologia digital**: como a natureza está transformando nossa tecnologia e nossas vidas. São Paulo: Berkeley Brasil, 2002.

BERNERS-LEE T.; LASSILA, O.; HENDLER, J. The semantic web. **Scientific American**, New York, v. 5, May 2001.

Disponível em: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 3 mar. 2009.

BLATTMANN, U.; SILVA, F. C. C. da. Colaboração e interação na web 2.0 e biblioteca 2.0. **Revista ACB:** Biblioteconomia em Santa Catarina, Florianópolis, v.12, n. 2, p. 191-215, jul./dez. 2007.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. 1997. 227 f. Tese (Doutorado). Centre for Telematics for Information Technology, University of Twence, Enschede, [1997]. Disponível em: <http:// http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf>. Acesso em: 11 fev. 2009.

BREWSTER, C.; CIRAVEGNA, F.; WILKS, Y. Background and foreground knowledge in dynamic ontology construction. In: ACM SIGIR WORKSHOP ON "SEMANTIC WEB" - SWIR, 2003, Toronto. **Report...** Disponível em: <http://www.sigir.org/forum/2003F/sigir03_ounis.pdf>. Acesso em: 7 jul. 2009.

BRICKLEY, D.; GUHA, R.V. **Resource Description Framework (RDF) SchemaSpecification 1.0**. 2000. Disponível em http://www.w3.org /TR/2000/CR-rdf-schema-20000327. Acesso em: out. 2008.

BUCKLAND, M. K. Description and search: Metadata as infrastructure. **Brazilian Journal of Information Science,** Marília, v. 0, n.0, p. 3-15, jul./dez. 2006. Disponível em <http://www.bjis.unesp.br>. Acesso em: 28 ago. 2009.

CAFÉ, L. et al. Repositórios institucionais: nova estratégia para publicação científica na Rede. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 26. 2003, Belo Horizonte. **Anais**... Belo Horizonte: INTERCOM, 2003. Disponível em: <http://www.intercom.org.br/papers/nacionais/2003/www/pdf/2003_ENDOCOM_TRABALHO_cafe.pdf >. Acesso em: 2 out. 2006.

CAMPOS, J.; SANTACHÈ, A.; TEIXEIRA, C. Visualização de modelos tridimensionais de sistemas de informações geográficas distribuídos baseados na WEB. In: BRAZILIAN WORKSHOP ON GEOINFORMATICS, 1999, Campinas. **Proceedings**... São José dos Campos: INPE, 1999. p. 50-58.

CAMPOS, M. L. de A.; GOMES, H. E. Taxonomia e classificação: o princípio de categorização. **DataGramaZero:** Revista de Ciência da Informação, Rio de Janeiro, v.9, n.4, ago. 2008. Disponível em: < http://www.datagramazero.org.br/ago08/Art_01.htm >. Acesso em: 13 abr. 2009.

CARDOSO, O. N. P. Recuperação de Informação. **InfoComp**, Lavras, v.2, n.1, 2000. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v2.1/olinda.pdf>. Acesso em: 21 nov. 2009.

CASTRO, F. F. de; SANTOS, P. L. V. A. C. MarcOnt Initiative: representação e descrição de recursos informacionais na web. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO- ENANCIB, 9., 2008, **Anais eletrônicos**... São Paulo: ANCIB, 2008.

CATARINO, M. E. **Integração das folksonomias nos metadados:** identificação de novos elementos como contributo para a descrição de recursos em repositórios. 233 f. 2009. Tese (Doutorado em Tecnologias e Sistemas de Informação) – Escola de Engenharia, Universidade do Minho, Guimarães, 2009.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? **IEEE Intelligent Systems**, IEEE Educational Activities Department, Piscataway, v. 14, n. 1, p. 20-26, 1999. ISSN 1541-1672.

DACONTA, M. C.; OBRST, L. J.; SMITH, K. T. **The Semantic Web:** a guide to the future of XML, Web Services, and Knowledge Management. Indiana: Wiley, 2003.

DCMI - DUBLIN CORE METADATA INITIATIVE. **Dublin Core Qualifiers**. 2008. Disponível em: <http://dublincore.org/documents/2008/ 01/14/dcmi-terms/>. Acesso em: 13 abr. 2009.

EVANS, P.; WURSTER, T. S. **Blown to bits: how the new economics of information transforms strategy.** Cambridge**:** Harvard Business School Press, 1999.

FARQUHAR, A.; FIKES, R.; RICE, J. **The ontolingua server: USA:** a tool for collaborative ontology construction. Duluth: Academic Press, 1997. p. 707-727.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. **Uma introdução sucinta à teoria dos grafos**. 2009. Disponível em: <http://www.ime.usp.br/~pf/teoriadosgrafos/>. Acesso em: 3 dez. 2009.

FERNEDA, E. Aplicando algoritmos genéticos na recuperação de informação, **DataGramaZero:** Revista de Ciência da Informação, Rio de Janeiro, v. 10, n. 1, fev. 2009. Disponível em: <http://www.dgz.org.br/fev09/F_I_aut.htm>. Acesso em: 21 out. 2009.

FERNEDA, E. **Recuperação da informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 147p. 2003. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicações e Artes, Universidade de São Paulo. Disponível em: <http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf>. Acesso em: 12 dez. 2008.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006.

FREITAS, F. L. G. **Ontologias e a Web Semântica**. Disponível em <http://www.inf.ufsc.br/~gauthier/EGC6006/material/Aula%203/Ontologia_Web_semantica%20Freitas.pdf>. Acesso em: 6 jun. 2008.

FUJITA, M. S. L. Organização e representação do conhecimento no Brasil: análise de aspectos conceituais e da produção científica do ENANCIB no período de 2005 a 2007. **Tendências da Pesquisa Brasileira em Ciência da Informação**, Brasília, v. 1, n. 1, 2008. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/4/13>. Acesso em: 2 fev. 2010.

GOMES, H. E. (Org.). **Manual de elaboração de tesauros monolíngues**. Brasília: Programa Nacional de Bibliotecas de Instituições de Ensino Superior, 1990.

GOMÉZ-PÉREZ, A. Ontological engineering: a state of the art, expert update, **British Computer Society**, London, v. 2, n.3, p.33-43, Autumn,1999.

GOODRICH, M. T.; TAMASSIA, R. **Estruturas de dados e algoritmos em java**. 2.ed. Porto Alegre: Bookman, 2002.

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

GREENBERG, J. The Semantic Web: more than a vision. **Bulletin for the American Society for Information Science and Technology,** Silver Spring, v. 29, n.4, p.6-7, Apr./May, 2003.

GRUBER, T. R. A translation approach to portable ontology specifications. **Technical Report KSL92-71**. Stanford: Knowledge Systems Laboratory. Stanford University, 1993. Disponível em: <http://www-ksl.stanford.edu/KSL_Abstracts/KSL- 92-71.html>. Acesso em: 15 fev. 2009.

GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing.** Padova. 1992. (Stanford University). Disponível em: < http://ksl.stanford.edu/KSL_Abstracts/KSL-93-04.html>. Acesso em: 15 fev. 2009.

GRUBER, T. R. **What is an ontology?** 1996. Disponível em: <http://ksl.stanford.edu/kst/what-is-an-ontology.html>. Acesso em: 29 abr. 2009.

GUARINO, N. Formal ontology and information systems. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS - FOIS'98, 1998, Trento. **Proceedings...** Amsterdam: IOS Press, 1998. p. 3-15.

GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: towards a terminological clarification. MARS, N. J. I. **Towards very large knowledge bases:** knowledge building and knowledge sharing. Amsterdam: IOS Press, 1995. p. 25-32. Disponível em: <http://www.csee.umbc.edu/771/papers/KBKS95.pdf.Z>. Acesso em: 8 fev. 2009.

GUIZZARDI, G**. Uma abordagem metodológica de desenvolvimento para e com reuso, baseada em ontologias formais de domínio**. 148 f. 2000. Dissertação (Mestrado em Informática) – Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2000.

GUY, M.; TONKIN, E. Folksonomies: tidying up tags? **D-Lib Magazine**, Reston, v.12, n.1, Jan. 2006. Disponível em: <http://wwww.dlib.org/dlib/ january06/guy/01guy.html>. Acesso em: 13 fev. 2009.

HARMELEN, F. V; MCGUINNESS, D. L. **OWL Web Ontology Language overview**. 2004. Disponível em <http://www.w3.org/TR/2004/REC-owl-features-20040210/> Acesso em: 5 jan. 2009.

HAYKIN, S. **Redes neurais**: princípios e prática. Porto Alegre: Bookman, 2001.

HORROCKS, I. et al. **DAML+OIL**. Disponível em: <http://www.daml.org/2001/03/daml+oil-index>. Acesso em: 1 jun. 2001.

IANNELLA, R.;WAUGH, A. **Metadata**: enabling the internet. CAUSE97, Melbourne, Apr. 1997. Disponível em: < http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=/published/emeraldfulltextarticle/pdf/2380200313_ref.html>. Acesso em: 22 nov. 2008.

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

JACOBSON, I.; BOOCH, G.; RUMBAUGH, J. **The unified software development process**. Reading: Addison-Wesley, 1999.

KALBACH, J. **Designing web navigation**: optimizing the user experience. Sebastopol: O'Reilly Media, 2007.

KOBASHI, N. Y. **Vocabulário controlado**: estrutura e utilização. 2008. (Mapeamento da oferta de capacitação nas escolas de governo). Disponível em:
<http://www2.enap.gov.br/rede_escolas/arquivos/vocabulario_controlado.pdf>. Acesso em: 2 dez. 2009.

KOOHANG, A. **Learning objects and instructional design**. Santa Rosa: Informing Science, 2007.

KURAMOTO, H. Informação científica: proposta de um novo modelo para o Brasil. **Ciência da Informação**, Brasília, v.35, n. 2, p. 91-102, maio/ago. 2006.

LAGOZE, C. The warwick framework: a container architecture for diverse sets of metadata. **D-Lib**, Arlington, July/Aug. 1996. Disponível em: <http://dlib.org/dlib/july96/lagoze/07lagoze.html>. Acesso em: 5 maio 2008.

LANCASTER, F. W. **Information retrieval systems**. New York: John Wiley, 1968.

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

LARMAN, C. **Utilizando UML e padrões**: uma introdução à análise e ao projeto orientados a objetos e ao desenvolvimento iterativo. 3. ed. São Paulo: Bookman, 2007.

LASSILA, O. Resource Description Framework (RDF) model and syntax **specification 1.0**. 1999. Disponível em: <http://www.w3c.org/TR/ REC-rdf-syntax>. Acesso em: 2 fev. 2009.

LEVACOV, M. Bibliotecas virtuais: (r)evolução? **Ciência da Informação**, Brasília, v.26, n.2, p.125-135, 1997.

LÉVY, P. **A inteligência coletiva**: por uma antropologia do ciberespaço. 2. ed. São Paulo: Loyola, 1999.

LIMA, V. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 14, n. 1, p. 131-151, 2009. Disponível em: <http://www.eci.ufmg.br/pcionline/index.php/pci/article/viewFile/729/543>. Acesso em: 3 dez. 2009.

LOPEZ YEPEZ, J. (Ed.). Fundamentos de informação e documentação. Madrid: EUDEMA, 1989. Recensão de: SILVA, L. A. G. da. **Ciência da Informação**, Brasília, v. 20, n.1, p. 95-97, jan./jun. 2001.

MAEDCHE, A.; STAAB, S. Semi-automatic Engineering of Ontologies from Text. In: **Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering**. 2000.

MARCONDES, C. H. Metadados: descrição e recuperação de informação na web. In: MARCONDES, C. H. et al. (Orgs.). **Bibliotecas digitais**: saberes e práticas. Salvador : Ed.UFBA; Brasília : IBICT, 2005. p. 97-114.

MARLOW, C. et al. **Position paper, tagging, taxonomy, Flickr, article, toRead**. Disponível em: <http://www.danah.org/papers/WWW2006.pdf>. Acesso em: 29 out. 2009.

MATHEUS, R. F.; SILVA, A. B. O. Fundamentação básica para análise de redes sociais: conceitos, metodologia e modelagem matemática. In: POBLACIÓN, D. A.; MUGNAINI, R.; RAMOS, L. M. S. V. C. **Redes sociais e colaborativas em informação científica**. São Paulo: Angellara, 2009. cap. 7, p. 239-287.

MAYRINK, D. F.; LADEIRA, A. P. Utilização de processamento automático de textos na construção de ontologias: um estudo de caso para a classificação de diagnósticos. In: CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE, 11., 2008, Campos do Jordão. **Anais...** , 2008. São Paulo: Sociedade Brasileira de Informática em Saúde, 2008. (CD-ROM). Disponível em: <http://www.sbis.org.br/cbis11/anais.htm>. Acesso em:14 dez. 2009.

MÉNDEZ, E.; BRAVO, A.; LÓPEZ, L. M. Microformatos: web 2.0 para Dublin Core. **El profesional de la información**, Barcelona, v. 16, n. 2, p. 107-113, marzo/abr. 2007.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, Washington, v. 2, n. 1, p.20-32. 1951.

MOREIRA, M. P.; MOURA, M. A. Construindo tesauros a partir de tesauros existentes: a experiência do TCI - tesauro em Ciência da Informação. **DataGramaZero:** Revista de Ciência da Informação, Rio de Janeiro, v. 7, n. 4, ago. 2006. Disponível em: < http://www.dgz.org.br/ago06/F_I_aut.htm>. Acesso em: 3 dez. 2009.

MOREIRA. A.; ALVARENGA, L.; OLIVEIRA, A. P. O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. **DataGramaZero**: Revista de Ciência da Informação, v.5, n. 6, dez. 2004. Disponível em: <http://dgz.org.br/dez04/Ind_art.htm>. Acesso em: 3 dez. 2009.

MORENO, F. P.; LEITE, F. C. L.; MÁRDERO ARELLANO, M. A. Acesso livre a publicações e repositórios digitais em Ciência da Informação no Brasil. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 1, p. 82-94, jan./abr. 2006.

NIC.BR. **Pesquisa sobre o uso das Tecnologias da Informação e da Comunicação no Brasil**. 2008. Disponível em <http://www.cetic.br /usuarios/index.htm> Acesso em: 6 jun. 2009.

NOVELLO, T. C. **Ontologias:** sistemas baseados em conhecimento e modelos de banco de dados. Universidade Federal do Rio Grande do Sul, 2002. Disponível em: <http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/ artigo_taisa.pdf>. Acesso em: 5 maio 2009.

OAI. **The open archives initiative protocol for metadata harvesting**. 2004. Disponível em: <http://www.openarchives.org/OAI/ openarchivesprotocol.html>. Acesso em: 18 mar. 2008.

OLIVEIRA, E. F. T. de; GRACIO, M. C. C. Rede de colaboração científica no tema "estudos métricos": um estudo de co-autorias através dos periódicos do Scielo da área de Ciência da Informação. **Brazilian Journal of Information Science**, Marília, v. 2, n. 2, p. 35-49, jul./dez. 2008. Disponível em: <http://www.bjis.unesp.br/pt/>. Acesso em: 21 dez. 2009.

O'REILLY, T. **What is web 2.0**: design patterns and business models for the next generation of software. 30 Sept. 2005. Disponível em: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html# mememap>. Acesso em: fev. 2009.

PEREIRA, A. M.; SANTOS, P. L. V. A. da C. O uso estratégico das tecnologias em catalogação. **Cadernos da Faculdade de Filosofia e Ciências**, Marília, v. 7, n. 1/2, p. 121-131, 1998.

PRIMO, A. O aspecto relacional das interações na Web 2.0. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 29., 2006, Brasília. **Anais...**, 2006. Brasília: UNB, 2006.

RAMALHO, R. A. S**. Web Semântica**: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. 2006. 120f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2006.

RODRIGUES E. et al**. RepositoriUM – implementação do DSpace em português:** lições para o futuro e linhas de investigação. Disponível em: < https://repositorium. sdum.uminho.pt/handle/1822 /679>. Acesso em: 2 maio 2009.

RUPLEY, S. What's a Wiki?. **PC Magazine**, 05 Sept. 2003. Disponível em: <http://www.pcmag.com/article2/0,4149,1071705,00.asp>. Acesso em: 21 jul. 2009.

SALES, R. de; CAFÉ, L. Diferenças entre tesauros e ontologias. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.14, n.1, p.17-98, jan./ abr. 2009. Disponível em: <http://www.eci.ufmg.br/pcionline/index.php/pci/article/view/646/541>. Acesso em: 3 jan. 2010.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, Oxford v. 24, n. 5, p. 513 – 523, 1988.

SALTON, G.; LESK. M. E. Computer evaluation of indexing and text processing. **Journal of the ACM**, New York, v. 15, n.1, p.8 – 36, Jan. 1968.

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

SANTOS, P. L. V. A. da C.; ALVES, R. C. V. Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0. **DataGramaZero**: Revista de Ciência da Informação, v.10, n. 6, dez. 2009. Disponível em: <http://www.datagramazero.org.br/dez09/Art_04.htm>. Acesso em: 3 dez. 2009.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYÃO, L. F. Padrões para bibliotecas digitais abertas e interoperáveis, **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, nº especial, p. 18-47, jan./jun. 2007. Disponível em <http://www.encontrosbibli.ufsc.br/bibesp/esp_06/bibesp_esp_06_sayao_esp_20071.pdf> Acesso em: 8 set. 2008.

SCHONS, C. H.; SILVA, F. C. C.; MOLOSSI, S. O uso de wikis na gestão do conhecimento nas organizações. **Biblios**: Revista de Bibliotecología e Ciências de la Información, Lima, v. 8, n. 27, p.1-10, enero/marzo 2007. Disponível em: http://redalyc.uaemex.mx/pdf/161/16102704.pdf. Acesso em: 11 dez. 2009.

SILVA, G. C.; LIMA, T. S. RDF e RDFS na infra-estrutura de suporte à websemântica. **Revista Eletrônica de Iniciação Científica**, Porto Alegre, v.2, n.2, mar. 2002. Sociedade Brasileira de Computação. Disponível em:< http://www.sbc.org.br/index.php?language=1&subject=101&content=magazine&option=content&id=3>. Acesso em: 22 fev. 2009.

SILVA, J. V. da; SILVA, S. R. P. da. Gerenciamento do vocabulário de tags do usuário em sistemas baseados em folksonomia. **Assembla**, p. 201-204, 2008. Disponível em: < http://www.assembla.com/spaces/folksonomy/documents/search?q=Gerenciamento+do+vocabul%C3%A1rio+de+tags+do+usu%C3%A1rio+em+sistemas+baseados+em+folksonomia.+&tag_name=&commit=Search>. Acesso em: 3 jan. 2010.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da**

Informação, Belo Horizonte, v.11 n.2, p. 161 -173, maio/ago. 2006. Disponível em: <www.eci.ufmg.br/pcionline/include/getdoc.php?id= 819&article=457&mode=pdf.> Acesso em: 13 dez. 2008.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a Ciência da Informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

SOUZA, T. B. et al. Metadados: catalogando dados na Internet. **Transinformação**, Campinas, v. 9, n.2, 1997, maio/ago. Disponível em: <http://puccamp.br/~biblio/tbsouza92.html>. Acesso em: 3 jan. 2009.

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

TÁLAMO, M. F. G. M.; KOBASHI, N. Y.; LARA, M. L. G. Contribuição da terminologia para a elaboração de tesauros. **Ciência da Informação**, Brasília, v.21, n.3, p.197-200, set./dez. 1992.

THE WEB STANDARDS PROJECT. **Web standards project**, 2009. Disponível em: < http://www.webstandards.org/>. Acesso em: 6 jun. 2009.

VIANA, C. L. M.; MÁRDERO ARELLANO, M. Á.; SHINTAKU, M. Repositórios institucionais em ciência e tecnologia: uma experiência de customização do DSpace. In: SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 3., 2005, São Paulo. **Proceedings...** São Paulo, 2005. p. 1-27. Disponível em <http://dici.ibict.br/archive/00000719/>. Acesso em: maio 2009

W3C CONSORTIUM. **Extensible Markup Language (XML).** 2009. Disponível em: <http://www.w3.org/XML>. Acesso em: 5 maio 2009.

WAL, T. V. **Folksonomy definition and wikipedia**. Disponível em: <http://www.vanderwal.net/random/entrysel.php?blog=1750>. Acesso em: 2 ago. 2009.

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

WIKIPEDIA. **O que a Wikipedia não é?** Disponível em < http://pt.wikipedia.org/wiki/Wikipedia>. Acesso em: 30 jul. 2009.

ZINS, C. et al. Knowledge Map of Information Science: Implications for the Future of the Field. **Brazilian Journal of Information Science,** Marília, v.1,

n.1, p.3-32, jan./jun. 2007. Disponível em: <http://www.bjis.unesp.br>. Acesso em: 2 ago. 2009.