Universidade Federal de Santa Catarina
Programa de Pós-Graduação em
Engenharia de Produção

Nikolai Dimitrii Braga de Albuquerque

UMA ARQUITETURA PARA O COMPARTILHAMENTO DO CONHECIMENTO EM BIBLIOTECAS DIGITAIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Vinícius Medina Kern, Dr.

Florianópolis
2003

Nikolai Dimitrii Braga de Albuquerque

UMA ARQUITETURA PARA O COMPARTILHAMENTO DO CONHECIMENTO EM BIBLIOTECAS DIGITAIS

Esta dissertação foi julgada e aprovada para obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 30 de setembro de 2003

_____

Prof. Edson Pacheco Paladini, Dr.
Coordenador do Programa

BANCA EXAMINADORA

_____

Prof. Vinícius Medina Kern, Dr.
Orientador
Universidade Federal de Santa Catarina

_____

Prof. Roberto C. S. Pacheco, Dr.
Universidade Federal de Santa Catarina

_____

Prof. José Leomar Todesco, Dr.
Universidade Federal de Santa Catarina

À Juliana, minha esposa, pela paciência e companheirismo.

À Daniela, minha filha, pela alegria em nossas vidas.

A meus irmãos, Andrik Dimitrii, Alexei Dimitrii e Farah Diba pelos momentos ausentes.

A meus pais, Ubirajara e Lêda pela dedicação.

AGRADECIMENTOS

Ao Grupo Stela por ter proporcionado o ambiente favorável ao desenvolvimento dessa dissertação, em particular ao Roberto Pacheco pelas experiências compartilhadas ao longo desses quatros anos de trabalho. Também sou muito grato ao professor Vinícius Kern que me orientado nessa etapa de pesquisa. Não poderia deixar de mencionar os amigos: André Castoldi, Cid Raulino, Turíbio, Rogério, Humberto Ferro, Lucas, Ricardo, Giovanni, Marcos Odaguiri, Rodrigo, Rosangela, Sandro Kerber, João Paulo, Tite e a família do Grupo Stela.

Em especial, gostaria de agradecer ao meu irmão Andrik Dimitrii pelo companheirismo em todo o desenvolvimento desse trabalho.

“O maior dos segredos é saber como reduzir a força da inveja.”

Cardinal de Retz

RESUMO

ALBUQUERQUE, Nikolai Dimitrii Braga de. **Uma arquitetura para o compartilhamento do conhecimento em bibliotecas digitais**. Florianópolis, 2003. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, 2003.

A World Wide Web é uma grande fonte de disseminação de informação, com destaque as bibliotecas digitais. O compartilhamento de recursos distribuídos, autônomos, heterogêneos e disponibilizados sem a mínima padronização gera a problemática da recuperação de uma elevada quantidade de informações que, na maioria das vezes, não atendem às necessidades dos usuários. Essa realidade está tornando-se um desafio para a comunidade científica que busca a integração, intercâmbio e entendimento semântico sobre essas informações. Tal integração engloba: o gerenciamento dos recursos, envolvendo a avaliação do conteúdo e de seus relacionamentos; e a padronização dos recursos, através da descrição de suas propriedades e implementação de mecanismos que dêem suporte à descoberta e recuperação dos mesmos. Várias iniciativas, como as desenvolvidas pelo World Wide Web Consortium, buscam por intermédio da criação de padrões, arquiteturas de metadados, serviços de inferência e ontologias, dentre outras, a melhor forma de tornar a informação também compreensível pelas máquinas. Este trabalho, propõe o desenvolvimento de uma arquitetura de integração semântica de bibliotecas digitais baseada no uso intensivo de padrões de metadados na forma de ontologias. Essas ontologias são armazenados em um repositório onde relações semânticas poderão ser estabelecidas, permitindo assim a interoperabilidade de repositórios de dados heterogêneos. A partir das tecnologias abordadas e da representação explícita da semântica do dado, aliados à teoria de ontologias, é possível oferecer serviços com um maior nível de qualidade, tornando a Web uma ferramenta capaz de tecer uma rede semântica de informações e ou conhecimentos.

Palavras-chave: compartilhamento do conhecimento, biblioteca digital, web semântica.

ABSTRACT

ALBUQUERQUE, Nikolai Dimitrii Braga de. **Uma arquitetura para o compartilhamento do conhecimento em bibliotecas digitais.** Florianópolis, 2003. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, 2003.

The World Wide Web is a great source of information dissemination, with prominence, the digital libraries. The distributed resources sharing, autonomous, heterogeneous and available without any standardization generates the problematic of the recovery over great amount of information that, in almost all the times, it doesn´t minister users necessities. his reality has becoming a challenge for the scientific community that search to integration, interchange and semantic understanding about these informations. This integration is composed: resources management, involving the content evaluation of the and their relationships; and the resources standardization, through yours properties description and implementation of mechanisms that gives support to discovery and recovery of it. Several initiatives, like those developed by the World Wide Web Consortium, try through the creation of patterns, architectures of metadata, inference services and onthology, among others, the better way of making information understandable for the machines. This work proposes the developmente of a semantic integration of digital libraries based on the intensive use of patterns of etadata in the form of onthologies. hese onthologies are stored in a repository where semantic relations will be established, thus allowing the interoperationability of heterogeneous data repositories. From these mentioned technologies and the explicit representation of the datum´s semantic, allied to the theory of onthologies, it is possible to offer services with a higher level of quality, turning the web a tool capable of weaving a semantic net of informations and/or knowledge.

Keywords: Knowledge Sharing, Digital Library, Semantic Web

SUMÁRIO

LISTA DE FIGURAS

LISTA DE REDUÇÕES

ASCII	*American Standard Code for Information Interchange*
BDB	*Biblioteca Digital Brasileira*
BDTD	*Biblioteca Digital de Teses e Dissertações*
CASE	*Computer Aided Software Enginnering*
CERN	*European Organization for Nuclear Research*
CML	*Conceptual Modelling Language*
CORBA	*Common Object Request Broker Architecture*
CWM	*Common Warehouse Model*
DAML	*DARPA Agent Markup Language*
DARs	*Distributed Active Relationships*
DC	*Dublin Core*
DLF	*Digital Library Federation*
DLG	*Directed Labeled Graph*
DRM	*Digital-Right Management*
DW	*Data Warehouse*
EDI	*Electronic Data Interchange*
FGDC	*Federal Geographic Data Committee*
GIGO	*Garbage in, Garbage out*
HTML	*Hipertext Markup Language*
IBICT	*Instituto Brasileiro de Informação Ciência e Tecnologia*
IDL	*Interface Definition Language*
LMPL	*Linguagem de Marcação da Plataforma Lattes*
KIF	*Knowledge Interchange Format*
MARC	*MAchine-Readable Catalogue*
MCF	*Meta Content Framework*
MDC	*Meta Data Coalition*
MOF	*Meta Object Facility*
MTD-BR	*Metadados de Teses e Dissertações Brasileiras*
OIM	*Open Information Model*
OKBC	*Open Knowledge Base Connectivity*
OMG	*Object Management Group*
OWL	*Ontology Web Language*
PICS	*Platform for Internet Content Selection*
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
SciELO	*Scientific Electronic Library Online*
SGML	*Standard Generalized Markup Language*
SQL	*Structured Query Language*

TOVE	*Toronto Virtual Enterprise*
UDK	*Umwelt-Datenkatalog*
UML	*Unified Modeling Language*
URL	*Uniform Resource Locator*
VSAM	*Virtual Storage Acess Method*
W3C	*World Wide Web Consortium*
WSDL	*Web Service Description Language*
WWW	*World Wide Web*
XMI	*XML Meta Interchange*
XML	*eXtensible Markup Language*
XSL	*eXtensible Style Language*
XSLT	*eXtensible StyleSheets Language Transformations*

1 INTRODUÇÃO

1.1 Apresentação

O crescimento da *World Wide Web* (WWW) é um acontecimento facilmente evidenciado principalmente na educação, onde a utilização de recursos na Web ganha cada vez mais importância devido ao fato de ser uma fonte natural e universal de pesquisa.

Esse crescimento acelerado exigiu uma demanda por serviços na Web num ritmo muito acelerado, acarretando na criação de uma infra-estrutura global e um conjunto de padrões para suportar a troca de documentos, um formato de apresentação para hipertexto (HTML) e técnicas para recuperação de informações. Porém, a alta heterogeneidade, autonomia e ampla distribuição dos dados na web sem uma padronização mínima tornaram a consulta dos dados pouco confiável, já que não se tem um esquema uniforme para se fazer consultas. As consultas são efetuadas geralmente através da navegação exaustiva, que dificilmente resulta em sucesso, devido ao grande volume de dados a ser consultado, ou através das buscas por palavras-chaves.

Devido a estas deficiências, a recuperação de informação na web vem se tornando um desafio para a comunidade científica, a qual, focou muitos estudos em soluções para os problemas emergidos com a consolidação e crescimento de sistemas web, principalmente no que tange à integração, intercâmbio e entendimento semântico das informações. Várias iniciativas concentram esforços na criação de padrões, modelos, linguagens, arquiteturas de metadados, dentre outros, para atender aos novos requisitos da web.

1.2 Justificativa

A WWW é hoje uma grande fonte de disseminação de informação nas principais áreas de conhecimento, mas o seu crescimento desordenado tem dificultado a localização, acesso, apresentação e manutenção das informações disponíveis. O compartilhamento de recursos distribuídos, autônomos, heterogêneos e disponibilizados sem a mínima padronização gera a problemática da recuperação da informação. A insatisfação dos usuários com relação à realização de consultas na Web está relacionada sobretudo com a recuperação de uma elevada quantidade de dados e ou informações que, na maioria das vezes, não atendem às necessidades dos usuários.

Atualmente as páginas web utilizam-se de mecanismos de representação embutidos em linguagens como HTML. No entanto, o conteúdo da informação é disponibilizado principalmente em linguagem natural, havendo portanto uma grande lacuna entre a informação disponibilizada e a recuperação realizada de forma automática por ferramentas.

Com base nessas deficiências, a Web se torna um desafio para a comunidade científica que busca sucesso na integração, intercâmbio e entendimento semântico sobre essas informações. Tal integração engloba: o gerenciamento dos recursos, envolvendo a avaliação do conteúdo e de seus relacionamentos; e a padronização dos recursos, através da descrição de suas propriedades e implementação de mecanismos que dêem suporte à descoberta e recuperação dos mesmos. Várias iniciativas, como as desenvolvidas pelo W3C, buscam por intermédio da criação de padrões, arquiteturas de

metadados, serviços de inferência e ontologias, dentre outras, a melhor forma de tornar a informação também compreensível pela máquina.

Os motivos que incentivaram o desenvolvimento desta pesquisa foi a dificuldade encontrada, na etapa de integração do Banco de Teses e Dissertações - BTD do PPGEP/UFSC com a Biblioteca Digital de Teses e Dissertações – BDTD. Projeto coordenado pelo IBICT e auxiliado por um consórcio de instituições das quais a UFSC faz parte.

1.3 Objetivo geral

O objetivo do presente trabalho é desenvolver uma arquitetura para integração semântica de bibliotecas digitais, contribuindo para a realização do compartilhamento do conhecimento.

1.4 Objetivos específicos

1. Especificar uma arquitetura para o gerenciamento e a integração semântica de bibliotecas digitais.

2. Apresentar um cenário de integração semântica em C&T utilizando os repositórios da Biblioteca Digital Teses e Dissertações Brasileira (BDTD) e a Biblioteca Digital da SciELO.

1.5 Metodologia

Quanto à natureza, este trabalho é considerado uma pesquisa aplicada, pois tem como objetivo a aplicação prática de um modelo dirigido à solução de problemas específicos, demonstrados na justificativa desta pesquisa. Quanto à abordagem do problema, este trabalho é considerado qualitativo. Ele está

baseado em um modelo de desenvolvimento em que não é necessária uma análise estatística para qualquer comprovação.

Do ponto de vista de seus objetivos, a pesquisa é exploratória, pois visa proporcionar uma maior relação com o problema através de um levantamento bibliográfico e de exemplos já existentes. Do ponto de vista dos procedimentos técnicos, esta pesquisa pode ser classificada como bibliográfica e de estudo de caso.

Para tratar o problema de pesquisa da dissertação, adotaram-se as seguintes etapas:

1. Estudo teórico sobre metadados, ontologias e web semântica no gerenciamento de conteúdos digitais.

2. Elaboração de uma arquitetura de integração semântica em bibliotecas digitais.

3. Aplicação da arquitetura de integração semântica utilizando a Biblioteca Digital Teses e Dissertações Brasileiras (BDTB) e a Biblioteca Digital da SciELO como fonte de dados.

1.6 Estrutura do Trabalho

O presente trabalho está estruturado em seis capítulos organizados de acordo com os temas relacionados ao desenvolvimento deste trabalho. No segundo capítulo é apresentada à definição e importância dos metadados no ciclo de vida e na integração dos sistemas informacionais. Também são apresentados padrões e arquiteturas de metadados que desempenham um

papel fundamental no processo de desenvolvimento de *software*. No terceiro capítulo são apresentadas definições sobre ontologias e a sua relação intrínseca com os metadados. Também são apresentados seus tipos e profundidades ontológicas e as metodologias e linguagens disponíveis no processo de desenvolvimento e integração. No quarto capítulo são apresentadas definições sobre a arquitetura da web semântica e a sua relação com a web atual. Também são apresentados a tecnologia dos web services e alguns protocolos recomendados pela W3C que desempenha um papel fundamental no processo de integração de repositórios de dados. No quinto capítulo é apresentada uma arquitetura de integração semântica e um protótipo de integração das bibliotecas digitais do IBICT e da SciELO. No sexto capítulo são apresentados as conclusões e trabalhos futuros.

2 METADADOS

O gerenciamento integrado de bases de dados heterogêneas não é uma tarefa trivial. A principal causa desta complexidade reside no fato de que a maior parte destes repositórios são produzidos de forma independente, usando tecnologias variadas de redes, sistemas operacionais e modelos de dados. Além disso, os dados normalmente encontram-se distribuído geograficamente o que contribui no aumento da complexidade, ainda mais se considerarmos distribuição no contexto da web. Prover o compartilhamento destas bases entre usuários de diferentes níveis tem sido o grande desafio da comunidade técnico-científica.

Algumas das características que dificultam o compartilhamento de dados segundo (GUNTHER, 1998) e (SIMON, 1998) são:

- Dados dinâmicos na sua conceitualização;
- A quantidade de dados a ser processada varia desde um grande volume de dados, como é o caso de processamento de imagens, até pequenos conjuntos de tabelas, como, por exemplo as tabelas contendo medidas de temperaturas coletadas de diversos sites e que podem apresentar uma diversidade de formatos;
- Os processos de captura dos dados, coleta, processamento e armazenamento são realizados geralmente através de diversas fontes de dados geograficamente distribuídas, o que torna o dado altamente distribuído e heterogêneo em termos de plataforma de *hardware* e *software*;

- Os dados podem estar organizados em uma variedade de modelos de dados (relacional, hierárquico, orientado a objetos, arquivos textos, etc.);
- Os dados são encontrados sob diferentes tipos como: medidas numéricas, variáveis, séries temporais e seqüências, imagens, documentos, sons, dentre outros;
- Os dados tendem a serem autônomos à medida que são produzidos e disponibilizados por instituições e organizações que operam de forma independente.

O alto grau de heterogeneidade apresentado pelo dado, no formato, nos diferentes ambientes de hardware e software nos quais se encontram armazenados, aponta para a necessidade de um mecanismo de integração que permita o acesso e a analise de dados de múltiplas fontes, que por sua vez representam diferentes domínios, de uma forma amigável, fácil e precisa.

Metadado tem sido considerado um elemento fundamental no suporte a interoperabilidade de recursos que apresentam um alto grau de distribuição e heterogeneidade. Sendo definido como sendo dado sobre o dado. O metadado auxilia na padronização da descrição, do processamento e da integração de dados heterogêneos (GUNTHER, 1997).

2.1 Processo de integração de dados

Segundo (HASSELBRING, 2000), o processo de integração de padrões em ambientes complexos envolve três principais aspectos: heterogeneidade, autonomia e distribuição, descritos a seguir:

- **Heterogeneidade:** é um dos principais fatores que dificulta a tarefa de integração entre padrões, é observada em dois níveis:
    - **Nível técnico:** resulta de diferentes plataformas de hardware, sistemas operacionais, sistemas de gerenciamento de bancos de dados e linguagens de programação;
    - **Nível conceitual:** resulta das diferentes interpretações do significado de certos termos, dos diferentes modelos de dados empregados (modelo relacional, modelo orientado a objeto, modelo de redes etc), bem como dos diferentes processos de modelagem para os mesmos conceitos do mundo real. Estes diferentes processos de modelagem derivam as discrepâncias esquemáticas (KRISHNAMURTHY, 1991), problema típico da área de banco de dados, onde o dado de uma base de dados corresponde ao metadado em outras bases de dados. Conciliar tais discrepâncias não é uma tarefa trivial considerando-se que cada modelo contém a semântica dos fatos expressa de diferentes formas. Em ambientes heterogêneos é comum encontrarmos dados com o mesmo conteúdo semântico, porém com nomes diferentes (sinônimos). Igualmente comum é a presença de homônimos para expressar conceitos diferentes.
- **Autonomia:** considerando os ambientes no contexto da web, as fontes de dados geralmente operam de forma independente ou semi-independente. Como conseqüência, estas fontes podem mudar os esquemas a qualquer momento sem haver qualquer tipo de autorização

ou mesmo de notificação. Conciliar estas mudanças nos esquemas de forma que os esquemas resultantes do processo de integração reflitam a realidade também não é uma tarefa trivial.

- **Distribuição:** os ambientes Web caracterizam-se por serem altamente distribuídos em termos de suas fontes de dados. Assim, é comum a distribuição entre microcomputadores, mainframes, servidores de redes locais e a própria Internet. Lidar com os diferentes protocolos de acesso aos dados representa um grande desafio no processo de integração.

Heterogeneidade semântica é reconhecida como um dos principais obstáculos no processo de interoperabilidade entre múltiplas fontes de dados, em especial no contexto de integração de esquemas (KENT, 1989), (KRISHNAMURTHY, 1991). Neste contexto, padrões de metadados estão sendo considerados cruciais na definição do significado da informação de modo que esta possa ser compartilhada por comunidades de diversas áreas do conhecimento.

2.2 O Papel do metadado no processo de integração

Atualmente existem diversos padrões de metadado voltados para domínios particulares do conhecimento, como por exemplo, a Linguagem de Marcação da Plataforma Lattes (LMPL) padrão que representa informações de C&T do Brasil e o Metadados de Teses e Dissertações Brasileiras (MTD-BR) padrão de teses e dissertações do IBICT. Entretanto, não existe um padrão comum capaz de atender toda e qualquer situação. Segundo (SIMON, 1998), não existe e

nunca existirá um padrão de metadado único devido à natureza heterogênea das aplicações.

Paralelo aos esforços para se obter uma definição de padrão de metadados, encontra-se a tendência em se descrever os conjuntos de dados e suas fontes com mais detalhes (qualidade do dado, informação histórica, etc.). Segundo (SIMON, 1997) é possível prover um melhor acesso aos dados através do uso de informação contextual, um texto livre associado ao dado.

O surgimento de diversos padrões de metadado originou um grande problema de incompatibilidade entre os padrões. O padrão *Dublin Core* (DC) (WEIBEL, 1999), padrão de referência no contexto de bibliotecas digitais, foi uma das primeiras tentativas de se gerar um padrão de metadado que fosse comum a todos os outros padrões. Apesar de ser um padrão aberto, ele não resolve o problema visto a natureza heterogênea de cada solução. É neste contexto que surgem as arquiteturas genéricas de metadados como solução para atingir interoperabilidade entre informações descritas em diferentes padrões de metadado.

2.3 Padrões e arquiteturas de metadados

A necessidade de se compartilhar grandes acervos sob uma perspectiva de um ambiente integrado tem levado a diversas iniciativas no contexto de padrões e arquiteturas de metadados por parte da comunidade técnico-científica. Neste sentido foram desenvolvidos padrões de metadados com finalidades específicas. Por exemplo, existem padrões que se preocupam somente com a representação de metadado, definindo que aspectos de um

recurso que devem ser descritos. Outros se preocupam com a troca de metadados, estabelecendo as interfaces necessárias. Um padrão para representação de metadado requer a completa descrição de um metamodelo com todos os seus elementos, seus conteúdos semânticos e os relacionamentos entre estes elementos. Este padrão deve ser totalmente independente de qualquer implementação específica. Um padrão para troca, por sua vez, é baseado em um único metamodelo e contém as definições de interface que especificam o metamodelo em linguagens do tipo *eXtensible Markup Language* (XML) e *Common Object Request Broker Architecture* (CORBA) *Interface Definition Language* (IDL).

As arquiteturas de metadados por sua vez, estabelecem mecanismos que permitem a codificação e o transporte de uma grande variedade de metadados desenvolvidos de forma independente, buscando assim garantir a interoperabilidade através do uso de convenções comuns a respeito da semântica, sintaxe e estrutura do metadado (IANELLA, 1998).

2.3.1 Padrões de Metadados

Padrões de metadados na área científica têm sido discutidos como mecanismos importantes para que agências governamentais, público em geral e a própria comunidade cientifica possam compartilhar seus acervos científicos. Nesta linha destacamos os padrões LMPL e MTD-BR, descritos a seguir.

- **LMPL:** segundo (Pacheco, 2001b), a Plataforma Lattes de sistemas de informação em ciência e tecnologia surgiu a partir da necessidade de integração de informações mantidas por CNPq, CAPES, Fapesp, Finep e outros sistemas do Ministério de Ciência e Tecnologia. Por ser uma

plataforma de C&T é indispensável sua integração com outros sistemas. Este esforço de compatibilização foi, de certa forma, precursor da LMLP, uma padrão de metadados para informações de C&T desenvolvida com o apoio de nove universidades (PUC-PR, UFBA, UFPE, UFRGS, UFRJ, UFRN, UFSC, Unicamp e USP). Seu desenvolvimento foi motivado pela necessidade de ampliação da Plataforma Lattes para atender a uma demanda das instituições de ensino e pesquisa.

- **MTD-BR:** é um padrão de metadados para teses e dissertações cuja principal finalidade é tornar disponível os meios para que a comunidade brasileira de C&T possa publicar seus trabalhos de forma rotineira, diretamente na rede, aumentando com isso sua visibilidade nacional e internacional, otimizando o fluxo da comunicação científica e reduzindo o ciclo de geração de novos conhecimentos. No capítulo 5 será apresentada uma definição mais detalhada sobre o MTD-BR bem como a utilização desse padrão no protótipo desenvolvido.

2.3.1.1 Interoperabilidade entre ferramentas

Nos últimos anos tem sido grande a preocupação entre os fabricantes de software com relação à interoperabilidade entre ferramentas. A troca de metadados entre estas é uma tarefa crítica, visto que estas ferramentas geralmente codificam e armazenam seus metadados de forma proprietária.

De modo a viabilizar o intercâmbio entre códigos gerados a partir de componentes de ferramentas distintas, alguns padrões foram criados. Dentre esses, os que mais se destacam são o OIM (OIM, 1999), o CWM (CWM, 2003) e o XMI (XMI, 1999).

OIM: é um padrão de metadados que surgiu da parceria de múltiplas empresas, algumas líderes de mercado, com o objetivo de prover suporte a interoperabilidade entre ferramentas de desenvolvimento, através da adoção de um modelo de informação compartilhado. Este padrão foi desenvolvido de forma a permitir o acompanhamento de todas as fases de desenvolvimento de um sistema de informação, desde a fase de análise até a fase de implantação. Adotado inicialmente como padrão pelo *Meta Data Coalition* (MDC), esse padrão baseia-se nos padrões de indústria *Unified Modeling Language* (UML), XML e *Structured Query Language* (SQL). Busca também prover o suporte a tecnologias de computação diversas, a exemplo de CASE, *intranet*, bancos de dados e *data warehouses*.

CWM: é um padrão de metadados cujo objetivo é permitir a integração de sistemas de *data warehouse* (DW), *e-business* e sistemas de negócios inteligentes em ambientes heterogêneos e distribuídos, através de uma representação e de um formato de troca de metadados. O padrão CWM é parte dos esforços do grupo *Object Management Group* (OMG), com o objetivo de prover um framework orientado a objeto e padronizado para aplicações distribuídas, visando dar suporte a reusabilidade, portabilidade e interoperabilidade entre componentes de DW. Adotado como um padrão OMG em junho de 2000. É baseado nos seguintes padrões OMG:

- UML como linguagem de modelagem padrão para organizar os tipos de metadados por assunto segundo categorias e funções num DW;
- *Meta Object Facility* (MOF), uma metalinguagem e um padrão de repositório de metadados;

– XMI, um padrão baseado em XML para troca de metadados entre ferramentas e repositórios orientados a objetos.

Os padrões OIM e CWM atuam no nível conceitual, aspectos de semântica, especificando metamodelos que podem ser vistos como os esquemas conceituais para os metadados, incorporando aspectos de aplicações específicas.

O padrão CWM, por sua vez, foi projetado para lidar somente com metadados no contexto de DW e provê um *framework* para representação e troca de metadados sobre as fontes de dados (origem e destino) envolvidas e os processos responsáveis pela criação e gerenciamento destas fontes.

XMI: é um padrão de metadado adotado pela OMG desde 1999, criado como uma iniciativa de fabricantes de *software* com o objetivo de prover interoperabilidade no contexto da OO entre ferramentas CASE, repositórios de metadados e ferramentas de desenvolvimento. Este intercâmbio é feito a partir de metadados armazenados em sistemas de arquivos tradicionais ou no formato de fluxo (*stream*) de dados baseados no padrão XML. O modelo de metadados utilizado tem como base o metamodelo MOF.

O padrão XMI, por sua vez, atua no nível físico, aspectos de sintaxe, preocupando-se em estabelecer um conjunto de regras capaz de gerar documentos XML a partir da especificação de modelos segundo o padrão MOF.

2.3.2 Arquiteturas de Metadados

Aspectos de interoperabilidade no nível semântico, sintático e estrutural são tratados pelas arquiteturas genéricas de forma a permitir que as informações,

descritas segundo os mais diferentes padrões, possam ser interpretadas e compartilhadas de forma adequada, evitando assim, a necessidade da unificação dos padrões de metadados (BARRETO, 1999).

O primeiro aspecto, interoperabilidade semântica, diz respeito à compreensão do significado de cada elemento componente dos diversos padrões de metadado, e pode ser alcançada através de duas abordagens (KERHERVÉ, 1997):

- ***bottom-up*:** onde a partir de diversos conjuntos de metadados desenvolvidos para atender as necessidades de uma determinada comunidade, deriva-se um único conjunto integrado e reduzido de forma que possa ser aplicado por esta comunidade. O padrão DC é um exemplo de uma abordagem *bottom-up* e, como já mencionado, não consegue solucionar o problema de se lidar com um grande número de padrões de metadados diferentes, uma vez que as soluções não convergem naturalmente a um denominador comum.

- ***top-down*:** onde a partir de um conjunto grande e bastante genérico de metadados especializa-se ou adapta-se para atender as necessidades de diversas comunidades e aplicações distintas. A *Resource Description Framework* (RDF) é um exemplo de arquitetura que emprega a abordagem *top-down*, que se mostra ser mais flexível e adaptável às necessidades das mais diferentes comunidades.

O segundo aspecto, interoperabilidade estrutural, refere-se ao modelo de dados empregado para definir a estrutura dos elementos componentes do padrão de metadado.

O modelo de dados pode variar de muito simples, utilizado para representar estruturas de metadados do tipo par (nome-elemento, valor-elemento), até muito complexo, utilizado para representar estruturas que envolvem hierarquia de classes e composição de classes.

O terceiro aspecto, interoperabilidade sintática, se refere à forma como os metadados são codificados para transferência. A sintaxe provê uma linguagem comum para representação das estruturas dos metadados. No contexto Web, XML é a linguagem que vem sendo utilizada para permitir a troca de metadados entre aplicações distintas.

Uma variedade de arquiteturas de metadados foi desenvolvida nos últimos anos, todas com um único objetivo: possibilitar a interoperabilidade entre provedores, catálogos e indexadores de modo a prover maior eficiência na descoberta de recursos de informação na Web. Nesse contexto, quatro arquiteturas podem ser citadas como contribuições importantes: *Warwick* (LAGOZE, 1996), *Meta Content Framework* (MCF) (GUHA, 1997), *Platform for Internet Content Selection* (PICS) e RDF. Dentre estas, a de maior destaque é a RDF, uma iniciativa da *World Wide Web Consortium* (W3C), que atualmente tornou-se a plataforma de desenvolvimento de aplicações na Web. Devido a sua grande importância no contexto da Web Semântica.

***Warwick*:** segundo (LAGOZE, 1996), essa arquitetura também é conhecida como arquitetura de recipientes, foi concebida para suportar qualquer conjunto de elementos de metadados. Os componentes básicos desta arquitetura, representados na Figura 2.1, são:

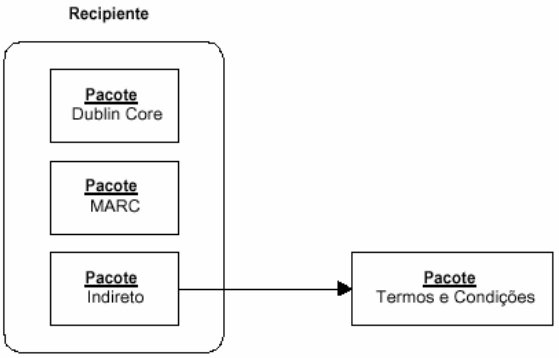

Figura 2-1: Exemplo da utilização da arquitetura Warwick

- **Recipiente:** representa a unidade básica para agregação de conjuntos de pacotes (metadados de determinado tipo).

- **Pacote:** representa uma estrutura de dados para armazenar metadados de um determinado tipo, dividida em três categorias:

    - **Pacote de conteúdo:** contém metadados de um determinado tipo (*MAchine-Readable Catalogue* (MARC), DC, e outros).

    - **Pacote indireto:** implementa uma referência a um objeto externo. Este objeto externo pode possuir seus próprios metadados e condições para acesso a algum recurso. Pacotes indiretos permitem o compartilhamento de objetos de metadados, à medida que o objeto alvo do pacote pode também ser indiretamente referenciado por outros recipientes.

    - **Pacote recipiente:** representa um pacote que também é um recipiente, armazenando ou servindo como meio de transporte para outros pacotes.

Uma extensão da arquitetura *Warwick* denominada de *Distributed Active Relationships* (DARs) foi proposta em (LAGOZE, 1996). Esta extensão tem por

objetivo definir um modelo para expressar relacionamentos entre recursos na Web, permitindo representar dado e metadado em objetos de biblioteca digital sem qualquer distinção evidente.

MCF: segundo (GUHA, 1997), essa é uma arquitetura aberta que foi concebida para ser utilizada na descrição da estrutura de *web sites* e de qualquer fonte de informação que possa estar contida na estrutura dos *sites*.

O modelo de dados da MCF possibilita descrever objetos, com seus atributos e relacionamentos com outros objetos, segundo tuplas de aridade n (geralmente 3), onde cada tupla corresponde a uma asserção que declara a existência de uma propriedade relacionada a um objeto. Este modelo é expresso segundo a estrutura de um *Directed Labeled Graph* (DLG), cujos elementos básicos, são representados na Figura 2.2.

- **Conjunto de Nós:** cada nó representa um objeto que pode ser um tipo primitivo (inteiro, caractere, etc.) ou uma entidade do mundo real (um documento, uma imagem, um mapa, uma pessoa, etc.).

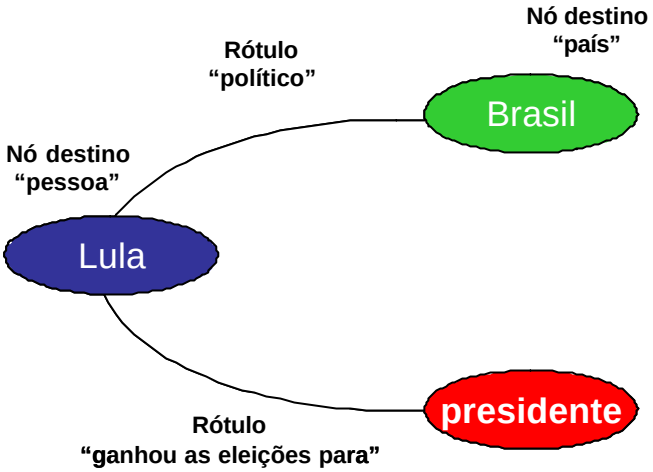

Figura 2.2 Modelo de dados da MCF

- **Conjunto de Rótulos:** cada rótulo representa um nome de propriedade que está associada ao objeto.

- **Conjunto de Arcos:** cada arco representa a associação entre os nós. A estrutura de um arco é uma tripla composta de um nó fonte, um nó destino e um rótulo, que representa a propriedade que vincula os nós.

O modelo de dados da MCF inclui um conjunto de tipos básicos que podem ser estendidos para acomodar novos tipos de dados. Além disso, também define um vocabulário básico que inclui um conjunto de termos comumente utilizados para descrição de conteúdos de documentos Web. Estes termos foram derivados de padrões já existentes como o padrão DC.

PICS: segundo (KRAUSKOPF, 1996) , a PICS é uma iniciativa do W3C, é um sistema para associação de metadado (denominado de PICS *labels*) ao conteúdo presente na Web. Inicialmente, foi concebido para ajudar pais e professores a controlarem o acesso das crianças a Internet, através de um formato comum para *labels*, de forma que qualquer *software* de seleção que

estivesse de acordo com o padrão PICS poderia processar qualquer *label* descrito segundo o padrão PICS. Atualmente, PICS tem sido considerada uma arquitetura concreta para o transporte de diferentes conjuntos de metadados associados a recursos de Internet, e o seu uso tem sido discutido em contextos como os de assinatura digital e privacidade.

RDF: o maior objetivo da arquitetura RDF é definir um mecanismo para descrever recursos não vinculados a um domínio específico de aplicação. Como resultado do trabalho em conjunto com várias comunidades, o RDF recebeu a influência de várias fontes diferentes. As principais influências vieram das comunidades de padronização da Web na forma de metadados *Hipertext Markup Language* (HTML) e PICS (PIC, 1996); de biblioteconomia; de estruturação de documentos na forma do *Standard Generalized Markup Language* (SGML) e XML; de representação do conhecimento e ainda de outras áreas de tecnologia que também contribuíram no projeto RDF: programação orientada a objetos, linguagem de modelagem e bancos de dados.

Na área de descoberta de recursos, por exemplo, a arquitetura RDF possibilita a implementação de mecanismos de pesquisa mais eficientes. Na área de catalogação a arquitetura RDF pode ser usada para descrever os recursos de informação em um sítio da Web, em uma biblioteca digital, etc. Na área de agentes inteligentes o RDF pode facilitar o intercâmbio de informações e o compartilhamento de conhecimento. A tecnologia RDF será abordada mais detalhadamente no capítulo 4, que trata a web semântica.

2.4 Considerações finais

Neste capítulo verificamos a necessidade da utilização dos metadados em todos os níveis organizacionais e principalmente computacionais, que são as bases das organizações. Como exemplos dessa necessidade, podemos citar a Internet que vem intensificando o fenômeno da “explosão” de documentos eletrônicos, ocasionando o aumento, considerado, do volume de informações disponíveis.

Diante dessa realidade, torna-se imprescindível o desenvolvimento de padrões de metadados que visem à descrição exata dos recursos de informação.

Nesse sentido, várias iniciativas estão sendo conduzidas com o propósito de discutir a questão e propor padrões de descrição de recursos de informação, como é o caso do *Dublin Core Metadata hitiative* e o padrão brasileiro de metadados para teses e dissertações (MTD-BR) desenvolvidos por um consorcio de instituições nacionais coordenadas pelo Instituto Brasileiro de Informação Ciência e Tecnologia (IBICT).

No próximo capítulo será descrito o papel das ontologias no contexto da padronização e relacionamento dos metadados utilizados nas organizações.

3 ONTOLOGIAS

A representação formal do conhecimento teve sua origem na Índia do primeiro milênio A.C. com o estudo da gramática de *Shastric Sanscrit*. No entanto, esta disciplina tem uma relação muito próxima dos trabalhos realizados na Grécia antiga, principalmente por Aristóteles (384-322 A.C.) nos campos da lógica, ciências naturais e filosofia metafísica (RUSSEL, 1995).

As primeiras discussões no campo da Inteligência Artificial focavam a questão da representação do ponto de vista do problema e não do conhecimento. Com a proliferação dos sistemas especialistas, a representação do conhecimento era feita com o objetivo claro de extrair o conhecimento do perito e formalizá-lo em uma base de conhecimento. Por outro lado, a maior parte dos esforços para incorporar conhecimento aos sistemas concentrava-se na construção de mecanismos uniformes e gerais de representação. Segundo (FALBO, 1998), a forma como o processo era conduzido influenciava diretamente alguns aspectos:

- Uma vez que a máquina de inferência era de propósito geral, a estratégia para resolver um problema era embutida como parte da base de conhecimento. Desta forma, era praticamente impossível separar os conhecimentos do domínio, da aplicação e da tarefa a ser realizada, tornando a reutilização do conhecimento praticamente inviável. O conhecimento do domínio não podia ser usado em outras aplicações dado que era adquirido para uma tarefa específica;

- O problema da reutilização era ainda agravado pelo modo no qual o conhecimento é associado e conseqüentemente disponibilizado por parte dos peritos. O conhecimento elucidado de especialistas em entrevistas é disponibilizado de forma bastante compilada através de heurísticas, o que dificulta a separação dos seus diversos tipos e praticamente inviabiliza o seu reuso;

- O uso do conhecimento do especialista como única fonte de conhecimento é, por si só, uma falha. A literatura técnica e sistemas existentes, desempenham papéis igualmente importantes, devendo ser utilizadas de forma complementar. Ao relegar estas fontes, a estratégia de transferência do conhecimento do especialista para o sistema não apenas tornava a tarefa de aquisição mais difícil, como também reforçava o problema da superficialidade.

Quando um sistema especialista era desenvolvido em um mesmo domínio, mas com o objetivo de realizar uma diferente tarefa, todo o processo de levantamento e codificação do conhecimento deveria ser refeito, expondo o processo a erros e inconsistências que já poderiam ter sido resolvidas, além de provocar perda de tempo, esforço e conseqüentemente recursos.

(CLANCEY, 1993) propõe a mudança desta perspectiva, argumentando que o foco da Engenharia de Conhecimento deve ser à modelagem de sistemas e não a tentativa de reproduzir a maneira como os especialistas raciocinam, defendendo a visão de que uma base de conhecimento deve ser vista como um produto de uma atividade de modelagem e não um repositório de conhecimento especializado. Desta forma, a modelagem passa a ser o aspecto

central da Engenharia de Conhecimento e a aquisição de conhecimento passa a ser essencialmente um processo construtivo, no qual o engenheiro de conhecimento usa todos os tipos de informação disponíveis e estabelece as decisões finais de modelagem. Dentro da comunidade de representação do conhecimento surgiu, então, um grupo de defensores da idéia de que o conhecimento embutido em uma determinada porção da realidade poderia (e deveria) ser representado em um nível de abstração tal que fosse independente e reutilizável ao longo de várias tarefas (GUARINO, 1997). Ao adotar este paradigma, esta comunidade entrou em um território anteriormente explorado unicamente por filósofos da ciência e da linguagem, fazendo com que, devido à imposição de sua disciplina, esta área fosse investigada de forma mais rápida e profunda do que quando era um domínio exclusivo da filosofia. Ao produto desta área inicialmente criada por Aristóteles com seu abrangente sistema de classificação, taxonomização e de representação do conhecimento de forma geral, chamamos hoje de Ontologias.

3.1. O que são Ontologias?

De acordo com o dicionário Webster (WOOLF, 1981), a palavra "ontologia" pode ser definida como uma teoria particular que diz respeito à natureza dos seres e das coisas em si. No sentido filosófico, o termo "ontologia" é referido como um sistema particular de categorias que versa sobre uma certa visão do mundo e pode ser visto como um sinônimo de metafísica. Seu propósito é classificar as entidades de uma porção da realidade, definindo seu vocabulário e as formulações canônicas de suas teorias (SMITH, 2000). Desta forma, este sistema não depende de uma linguagem particular. Por exemplo, uma

ontologia de Aristóteles é sempre a mesma, independente da linguagem usada para expressá-la. Por outro lado, para a comunidade das Ciências da Computação, o termo se refere a um artefato de engenharia, constituído de um vocabulário de termos organizados em uma taxonomia, suas definições e um conjunto de axiomas formais usados para criar novas relações e para restringir as suas interpretações segundo um sentido pretendido (NOY, 1997). Apesar da relação entre essas duas definições, com o intuito de resolver este impasse terminológico, em (GUARINO, 1998b), é proposto que a definição da comunidade de computação seja adotada para o termo "ontologia" e que para a definição filosófica seja dado o nome de conceituação.

Nesse trabalho, o termo ontologia é usado em concordância com a definição de (GUARINO, 1998b), ou seja, ontologias são tratadas como um artefato computacional composto de um vocabulário de conceitos, suas definições e suas possíveis propriedades, um modelo gráfico mostrando todas as possíveis relações entre os conceitos e um conjunto de axiomas formais que restringem a interpretação dos conceitos e relações, representando de maneira clara e não ambígua o conhecimento do domínio.

3.2. Tipos de Ontologias

Segundo Guarino (GUARINO, 1997, 1998b), com base em seu conteúdo as ontologias podem ser classificadas nas seguintes categorias:

- **ontologias genéricas:** expressam teorias básicas do mundo, de caráter bastante abstrato, aplicáveis a qualquer domínio, conhecimento de senso comum. Tipicamente, ontologias genéricas definem conceitos tais

como coisa, estado, evento, processo, ação, etc., com o intuito de serem especializados na definição de conceitos em uma ontologia de domínio;

- **ontologias de domínio:** expressam conceituações de domínios particulares, descrevendo o vocabulário relacionado a um domínio genérico, tal como medicina e direito. São construídas para serem utilizadas em um micro-mundo;

- **ontologias de tarefas:** expressam conceituações sobre a resolução de problemas, independentemente do domínio em que ocorram, isto é, descrevem o vocabulário relacionado a uma atividade ou tarefa genérica, tal como, diagnose ou vendas. O estudo de ontologias de tarefas é a vertente mais recente do estudo de ontologias. Sua principal motivação é facilitar a integração dos conhecimentos de tarefa e domínio em uma abordagem mais uniforme e consistente, tendo por base o uso de ontologias. Trabalhos nesta categoria incluem (CHANDRASEKARAN, 1997), (MUSEN, 1995);

- **ontologias de aplicação:** expressam conceitos dependentes das ontologias do domínio e das ontologias de tarefa. Estes conceitos freqüentemente correspondem a papéis desempenhados por entidades do domínio quando da realização de uma certa atividade;

- **ontologias de representação:** expressam os compromissos ontológicos embutidos em formalismos de representação de conhecimento. Um exemplo desta categoria é a ontologia de *frames*, utilizada em Ontolingua (GRUBER, 1992).

(GUARINO, 1998b) propõe que ontologias sejam construídas segundo seu nível de generalidade, como é mostrado na Figura 3.1. Os conceitos de uma ontologia de domínio ou de tarefa devem ser especializações dos termos introduzidos por uma ontologia genérica. Os conceitos de uma ontologia de aplicação, por sua vez, devem ser especializações dos termos das ontologias de domínio e de tarefa.

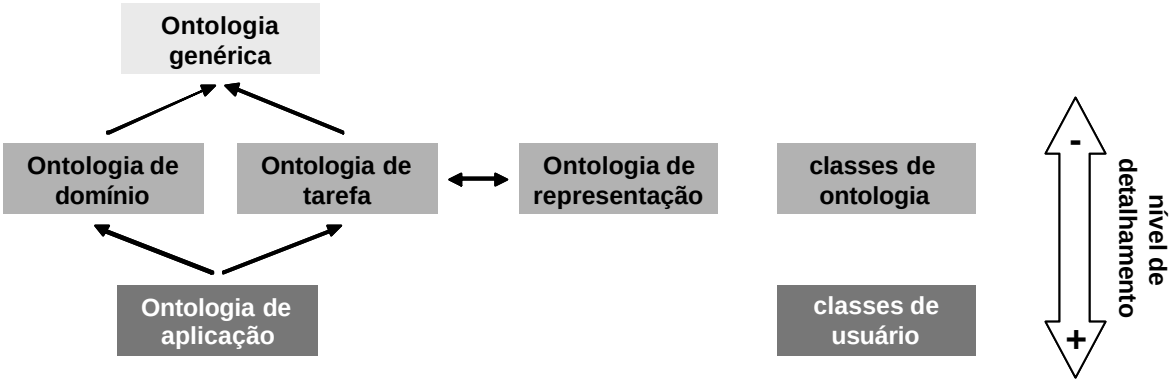

Figura 3.1 - Tipos de ontologias, segundo seu nível de dependência (Guarino, 1998)

3.3. Profundidade ontológica

Quanto à profundidade ontológica, (GUARINO, 1998a) define quatro níveis:

- **Vocabulário:** Em sua forma mais simples, uma ontologia é apenas um vocabulário. Nesse sentido, uma DTD ou um XML-*Schema* pode definir uma ontologia. Se diferentes parceiros combinarem com uma XML-*Schema*, eles também concordam com a ontologia definida através do *Schema*, pois os nomes de *tag* declarados nele definem um vocabulário comum. Quando, por exemplo, dois agentes concordam em compartilhar tal ontologia, eles podem trocar mensagens usando o vocabulário da ontologia;

- **Taxonomia:** o significado dos termos na taxonomia é estabelecido pela definição de relacionamentos entre eles. Um dos relacionamentos mais naturais é a classificação – o estabelecimento de relacionamentos entre objetos e classes, subclasses e classes pai. Esses sistemas são denominados taxonomia, sendo que suas relações são conhecidas como relacionamentos “é-um”. Esse tipo de ontologia normalmente é estabelecido por sistemas orientados a objetos. Muitas ontologias existentes são definidas usando-se apenas esses relacionamentos hierárquicos;

- **Sistema Relacional:** as ontologias também podem incluir relacionamentos não hierárquicos. Existem muitos relacionamentos possíveis entre os objetos além dos relacionamentos hierárquicos “é-um”. Tais relacionamentos são típicos nos diagramas de relacionamento de entidades e nos bancos de dados relacionais e, por conseguinte, cada esquema de banco de dados relacional define sua própria ontologia;

- **Teoria Axiomática:** além de escrever relacionamentos, as ontologias também podem impor restrições. As restrições são definidas como axiomas. Um axioma é uma afirmação lógica que não pode ser provada a partir de outras afirmações. Nos bancos de dados relacionais, as restrições são definidas por regras de integridade. Além disso, algumas linguagens orientadas a objeto incluem a capacidade de definir restrições (por exemplo, asserções em C++ e contratos em Eiffel).

3.4. Benefícios no uso de ontologias

De forma geral, ontologias constituem uma ferramenta poderosa para suportar a especificação e a implementação de sistemas computacionais de qualquer complexidade. Ao usar esta abordagem na fase de análise e especificação do domínio, podemos destacar três áreas:

- **Comunicação:** ontologias são ferramentas úteis para ajudar as pessoas a se comunicarem, sob várias formas, acerca de um determinado conhecimento. Em primeiro lugar, elas podem ajudar as pessoas a raciocinar e a entender o domínio do conhecimento e, portanto, atuam como uma referência para a obtenção do consenso numa comunidade profissional sobre o vocabulário técnico a ser usado nas suas interações. Além disso, ontologias constituem um excelente guia no processo de elicitação de conhecimento das diversas fontes;

- **Formalização:** devido à natureza formal da notação usada, a especificação do domínio elimina contradições e inconsistências envolvendo as restrições, resultando, portanto, em uma especificação não ambígua. Um outro ponto a ser destacado é que, já que uma notação formal é usada, a especificação formalizada pode ser automaticamente verificada e validada, se um provador automático de teoremas existe para aquela notação. Com um mecanismo de inferência, é também possível derivar novos conhecimentos de forma automática, a partir da base de conhecimento já presente na ontologia. Por fim, esta característica torna possível a obtenção de um processo de geração de

infra-estruturas computacionais de maneira sistemática e idealmente automática;

- **Representação do conhecimento e reuso:** A ontologia forma um vocabulário de consenso e representa o conhecimento do domínio de forma explícita no seu mais alto nível de abstração, possuindo um potencial enorme de reuso. O conhecimento formalizado na camada de domínio pode ser especializado em diferentes aplicações, servindo diferentes propósitos, por diferentes equipes de desenvolvimento, em diferentes pontos do tempo.

3.5. Problemas no uso de ontologias

Apesar de todas as vantagens citadas, o uso de ontologias também apresenta alguns problemas. (O'LEARY, 1997), identifica os seguintes:

- A escolha de uma ontologia é um processo político, já que nenhuma ontologia pode ser totalmente adequada a todos os indivíduos ou grupos;

- Ontologias não são necessariamente estacionárias, i.e., necessitam evoluir. Poucos trabalhos têm enfocado a evolução de ontologias;

- Estender ontologias não é um processo direto. Ontologias são, geralmente, estruturadas de maneira precisa e, como resultado, são particularmente vulneráveis a questões de extensão, dado o forte relacionamento entre complexidade e precisão das definições;

- A noção de bibliotecas de ontologias sugere uma relativa independência entre diferentes ontologias. A interface entre elas constitui, portanto, um impedimento, especialmente porque cada uma delas é desenvolvida no contexto de um processo político. Ontologias desenvolvidas independentemente podem não se integrar efetivamente com outras por vários motivos, desde similaridade de vocabulário até visões conflitantes do mundo (FALBO, 1998). O pior problema, no entanto, é do ponto de vista metodológico. Apesar de uma grande quantidade de ontologias já ter sido desenvolvida por diferentes grupos, sob diferentes abordagens e usando diferentes métodos e técnicas, poucos trabalhos foram publicados sobre como proceder, mostrando as práticas, critérios de projeto, atividades, métodos e ferramentas usadas para sua construção. A conseqüência é clara: a ausência de atividades padronizadas, ciclos de vida e métodos sistemáticos, assim como de um conjunto de critérios de qualidade, técnicas e ferramentas, expõem o desenvolvimento de ontologias aos mesmos problemas da engenharia de software, ou seja, a sua realização é conduzida de forma artística e não como uma atividade de engenharia (FALBO, 1998).

3.6. Construção de ontologias

Algumas propostas de metodologias para construção de ontologias têm sido apresentadas na literatura nos últimos anos, como por exemplo a "metodologia inicial" apresentada por (USCHOLD, 1995), *METHONTOLOGY* (FÉRNANDEZ, 1997) e a apresentada no contexto do projeto *Toronto Virtual Enterprise* (TOVE) (USCHOLD, 1996). Apesar disso, os modelos apresentados ainda não

demonstram um processo suficientemente estruturado a ponto de suportar a construção de ontologias como uma verdadeira disciplina de engenharia. Nessa seção, é apresentada uma abordagem sistemática para construção de ontologias, assim como descrita por (FALBO, 1998). Esta abordagem, além de unir as principais características das metodologias citadas, discute as várias atividades do processo de construção de ontologias, apresentando algumas orientações de como proceder na sua realização. É importante ressaltar que, devido à complexidade envolvida nas atividades que compõem este ciclo, a área de engenharia de ontologias urge pelo surgimento de ferramentas CASE que possam realizar a automatização (ou pelo menos semi-automatização) do processo.

- **Identificação de propósito e especificação de requisitos:** A primeira atividade a ser realizada no processo de construção de uma ontologia é identificar claramente o seu propósito e os usos esperados para ela (FALBO, 1998), i.e., a competência da ontologia. A competência de uma representação diz respeito à cobertura de questões que essa representação pode responder ou de tarefas que ela pode suportar. Ao se estabelecer à competência, temos um meio eficaz de delimitar o que é relevante para a ontologia e o que não é. É útil, também, identificar potenciais usuários e os cenários que motivaram o desenvolvimento da ontologia em questão. Uma vez definido o propósito, deve-se especificar os requisitos da ontologia. Esses devem contemplar os usos projetados para a ontologia e podem ser expressos em termos de questões de competência: as questões que a ontologia deve ser capaz de responder (USCHOLD, 1996). Ao se especificar um relacionamento entre as

questões de competência e os cenários de motivação, está se dando uma justificativa para a ontologia e, mais importante, se está provendo um mecanismo para sua avaliação. Uma analogia pode ser feita entre o papel que as questões de competência desempenham para a engenharia de ontologias, comparando-o ao dos modelos de casos de uso no contexto da engenharia de software orientada a objetos. Ambas as técnicas norteiam todo o processo de desenvolvimento, auxiliando deste a atividade de especificação de requisitos até a atividade de avaliação;

- **Captura da ontologia:** Esta é, sem dúvida, a etapa mais importante no desenvolvimento de uma ontologia. O objetivo é capturar a conceituação do universo de discurso, com base na competência da ontologia. Os conceitos e relações relevantes devem ser identificados e organizados. Um modelo utilizando uma linguagem gráfica pode ser de grande utilidade para facilitar a comunicação com os especialistas do domínio. Este modelo deve ser acompanhado de um vocabulário de termos (FALBO, 1998). Conceitos primitivos, isto é, aqueles que não são passíveis de uma definição em termos de outros conceitos da ontologia, devem ser definidos utilizando linguagem natural e exemplos, tomando o devido cuidado para se evitar ambigüidades e inconsistências. A escolha dos termos a serem usados para referenciar as categorias de conhecimento deve ser feita cuidadosamente, evitando-se termos com interpretação duvidosa. Conceitos passíveis de descrição em termos de outros conceitos, devem ser definidos com referências claras a estes, com o objetivo de facilitar a formalização (FALBO, 1998). Deve-se,

ainda, construir taxonomias, organizando categorias e sub-categorias interconectadas do conhecimento do domínio de interesse. Os conceitos e relações formam a base da ontologia. Mas uma característica essencial de ontologias é a definição de axiomas. Simplesmente propor uma taxonomia ou um conjunto de termos básicos, não constitui uma ontologia. Axiomas devem ser providos para definir a semântica dos termos. Os axiomas especificam definições de termos na ontologia e restrições sobre sua interpretação. Neste momento, não há necessidade de se escrever axiomas formais mas, ao contrário, estes devem ser descritos em linguagem natural, refletindo simplesmente as restrições existentes sobre o universo de discurso. Os axiomas em uma ontologia podem apresentar duas formas e propósitos diferentes: axiomas de derivação e axiomas de consolidação. Axiomas de derivação são aqueles que permitem explicitar informações a partir do conhecimento previamente existente. Assim, são meios para a dedução e representam conseqüências lógicas neste processo. Axiomas de consolidação, por sua vez, não são utilizados para derivar informação, mas apenas para descrever a coerência das informações existentes. Neste sentido, não representam conseqüências lógicas. Tipicamente, os axiomas de consolidação definem condicionantes para o estabelecimento de uma relação ou para a definição de um objeto como instância de um conceito (FALBO, 1998). Os axiomas de derivação podem ter origem no significado dos conceitos e relações da ontologia ou na forma como são estruturados. Quando axiomas são descritos para mostrar restrições impostas pela forma de estruturação dos conceitos, eles são ditos

axiomas epistemológicos. Quando descrevem restrições de significação impostas no domínio, são ditos axiomas ontológicos (FALBO, 1998). Esta classificação quanto à natureza dos axiomas é uma boa diretriz para guiar a definição dos axiomas de uma ontologia, ou seja, devemos estar atentos para capturar axiomas que considerem a estruturação dos conceitos e relações (os axiomas epistemológicos), seus significados e restrições (os axiomas ontológicos) e as leis de integridade que os regem (os axiomas de consolidação) (FALBO, 1998). O processo de definição de axiomas é, talvez, o aspecto mais difícil na construção de ontologias. Entretanto, esse processo pode e deve ser fortemente guiado pelas questões de competência. Os axiomas devem ser necessários e suficientes para expressar as questões de competência e para caracterizar suas soluções. Além disso, qualquer solução para uma questão de competência deve ser descrita pelos axiomas da ontologia e deve ser consistente com eles. Se os axiomas propostos não forem suficientes para esse propósito, então conceitos, relações ou axiomas adicionais devem ser introduzidos na ontologia. Por outro lado, axiomas redundantes ou que não contribuem para responder a uma questão de competência devem ser eliminados (FALBO, 1998). Neste sentido, a captura de uma ontologia é um processo iterativo e fortemente ligado à avaliação (USCHOLD, 1996);

- **Formalização da ontologia:** Para a realização desta etapa, é necessário que um formalismo de representação das diversas categorias de conhecimento da ontologia seja escolhido. À primeira vista, qualquer linguagem de representação formal do conhecimento, ou mesmo

informal, poderia ser usada para representar ontologias (FALBO, 1998). Na prática, entretanto, apenas poucas linguagens têm sido usadas para este propósito, entre elas: lógica de primeira ordem, *Knowledge Interchange Format* (KIF) (GRUBER, 1992), Ontolingua (GRUBER, 1995), *Conceptual Modelling Language* (CML) (BREUKER, 1994) e *Description Logic* (RUSSEL, 1995). A validação de uma teoria sobre um universo de discurso é, sem dúvida, melhor realizada quando esta é descrita em uma linguagem formal, ou seja, uma linguagem fundamentada em um modelo matemático. Nesta linguagem, em contraste com a linguagem natural, têm-se símbolos não ambíguos e formulações exatas e, portanto, a clareza e a correção de uma dedução podem ser testadas com maior facilidade e precisão. Uma dedução em linguagem natural, geralmente, envolve pressuposições implícitas que entram desapercebidas no processo de dedução. O tratamento teórico de qualquer domínio consiste em propor sentenças sobre os objetos neste domínio (sentenças atribuindo certas propriedades e relações aos objetos em questão) e em estabelecer regras de acordo com as quais outras sentenças possam ser derivadas a partir das sentenças dadas. É importante ressaltar que todas essas linguagens possuem vantagens específicas e assumem compromissos ontológicos em níveis variados, e portanto a escolha de que linguagem usar depende diretamente do propósito da ontologia;

- **Integração com ontologias existentes:** Durante os processos de captura e/ou formalização, pode surgir à necessidade de integrar a ontologia em questão com outras já existentes, visando aproveitar

conceituações previamente estabelecidas. De fato, é uma boa prática desenvolver ontologias funcionais modulares, que sejam gerais e mais amplamente reutilizáveis, e, quando necessário, integrá-las, obtendo o resultado desejado (FALBO, 1998);

- **Avaliação:** A ontologia deve ser avaliada para verificar se satisfaz os requisitos estabelecidos na especificação. Esta etapa deve ser realizada em paralelo com as etapas de captura e formalização. (GRUBER, 1995) apresenta um conjunto de critérios para guiar tanto o desenvolvimento, quanto para avaliação da qualidade das ontologias construídas. Os principais critérios definidos são: clareza, coerência, extensibilidade e compromissos ontológicos mínimos. Em (USCHOLD, 1996), é defendido que, adicionalmente, as questões de competência devem ser usadas principalmente para avaliar a adequação da axiomatização realizada;

- **Documentação:** Todo o desenvolvimento da ontologia deve ser documentado, incluindo propósitos, requisitos e cenários de motivação, as descrições textuais da conceituação, a ontologia formal e os critérios de projeto adotados. Como foi dito anteriormente, assim como a avaliação, a documentação é considerada uma atividade guarda-chuva do processo, ou seja uma etapa que deve ocorrer durante todas as iterações do ciclo em paralelo com as demais. Os termos capturados na conceituação do universo de discurso devem ser descritos em um Dicionário de Termos, considerando dois princípios importantes:

- **Princípio do vocabulário mínimo**: indica o vocabulário utilizado na definição dos termos da ontologia. Este vocabulário deve ser o menor possível e não deve apresentar ambigüidades;

- **Princípio da auto-referência**: indica que a definição de um termo no Dicionário deve, sempre que possível, ser feita utilizando outros termos já mencionados. Com base neste princípio, o uso de hipertextos surge como uma abordagem para a documentação de ontologias. Esta tecnologia mostra-se adequada, tendo em vista que torna natural a definição de novos termos a partir de outros mais primitivos, permitindo navegação entre definições, exemplos e a formalização, incluindo os axiomas (FALBO, 1998).

3.7. Linguagens para a criação de ontologias

Para a construção de uma ontologia são utilizados os seguintes objetos:

- **Entidades:** descrevem conceitos (elementos de um domínio estudado) e providenciam representações lógicas;
- **Atributos:** descrevem as propriedades das entidades;
- **Relações:** descrevem as ligações entre os objetos no modelo (entidades e atributos);
- **Restrições:** condições que o projetista impõe sobre as entidades, atributos ou relações.

Diversas linguagens e mecanismos para a definição de ontologias foram criados nos últimos anos, a exemplo de: SHOE (LUKE, 2000), XOL (KARP,

1999); OIL (FENSEL, 2000) (HORROCKS, 2000), *DARPA Agent Markup Language* (DAML[1]), SontoDL (GRUTTER, 2001), dentre outros.

A principal característica dessas linguagens está na capacidade de representar ontologias em RDF/RDFS, arquitetura já consagrada pela W3C para interoperabilidade de informações na Web, essa tecnologia será apresentada no capítulo 4. A seguir serão apresentadas algumas dessas linguagens que tem merecido maior destaque na literatura.

SHOE (Simple HTML Ontology Extensions): é uma extensão da HTML (recentemente também utilizado em XML) que provê meios de incorporar conhecimento semântico de forma a ser entendido por máquina (robôs, agentes, etc.) ou outros documentos WWW. O objetivo é facilitar os mecanismos de busca.

A linguagem SHOE inclui um mecanismo de definição de ontologias, instâncias de dados em páginas Web e de classificação hierárquica de documentos HTML. Isto é feito a partir de categorias (classes) e regras que especifica relacionamentos e hierarquias entre instâncias, a partir de um conjunto de *tags* acrescidos ao HTML padrão.

XOL (XML-based Ontology Exchange Language): é uma linguagem de especificação e intercâmbio de ontologias, especificado em DTD ou XML *Schema*. Utiliza um modelo semântico baseado em *frames* denominado *Open Knowledge Base Connectivity* (OKBC). Um arquivo XOL consiste de um módulo cabeçalho de definição contendo: nome e versão que estabelece

[1] http://www.daml.org/about.html

metas-informação sobre a ontologia; classes e subclasses que estabelece hierarquias entre categorias de elementos; *slots* que estabelece propriedades aos elementos das classes e definições individuais que permite declarar nomes, descrições, informações sobre instância, valores, etc.

OIL (Ontology Inference Layer): é uma linguagem para a especificação de ontologias que reúne as seguintes características: provê primitivas de modelagem normalmente utilizadas em ontologias baseadas em *frames*; possui semântica bem definida, simples e clara baseada em descrição lógica; e apresenta suporte para dedução automática (BECHHOFER et al., 2000).

Uma ontologia OIL contém descrições para classes, relacionamentos (*slots*) e instâncias. Classes podem se relacionar com outras classes através de uma hierarquia (classes/subclasses) e através de relações binárias estabelecidas entre duas relações. Além disso, restrições de cardinalidade podem ser atribuídas aos relacionamentos.

A definição de uma ontologia em OIL é constituída de dois componentes: o primeiro, que descreve as características da ontologia (*ontology container*) utilizando-se de descritores do padrão DC; e o segundo (*ontology definitions*) que define o vocabulário particular daquela ontologia.

A linguagem OIL tem sido considerada pela W3C como uma linguagem de grande relevância no contexto atual de desenvolvimento de aplicações na Web.

DAML (DARP Agent Markup Language): é uma iniciativa da agência DARPA que está sendo desenvolvida como uma extensão da XML e RDF. A sua mais recente iniciativa é oriunda da combinação de DAML e OIL

(DAML+OIL, 2000), uma linguagem padrão para representação de ontologias e metadados pela W3C.

A combinação de DAML e OIL, denominada DAML+OIL, sofre muita influência do OIL original, embora não utilize o seu conceito original de *frames*. É constituído de uma coleção de classes, propriedades e objetos que são adicionados ao RDF e RDFS. Assim, declarações (*statements*) em DAML+OIL também são declarações RDF.

OWL (*Web Ontology Language*): é a linguagem de marcação semântica para representar e compartilhar ontologias na Web. É derivada da DAML+OIL e é construída sob a sintaxe RDF, disponibilizando outras características que se concentram basicamente em proporcionar maiores flexibilidades para propiciar que as ontologias possam ser definidas sob uma forma semântica cada vez mais rica em detalhes. Segundo (CASTOLDI, 2003), a OWL é dividida em três sub-linguagens:

- **OWL Lite:** suporta os usuários que precisam basicamente de uma hierarquia de classificação e funcionalidades de restrição básicas.
- **OWL DL:** atende aos usuários que desejam a máxima expressividade ao mesmo tempo em que seus sistemas inteligentes mantenham decidabilidade (todos os cálculos terminarão em um tempo finito) e completude (garantida de que todas as conclusões serão executadas) computacional.
- **OWL Full:** destina-se àqueles que desejam o máximo em poder de expressão e a liberdade sintática do RDF, sem garantias

computacionais. OWL Full permite que uma ontologia aumente o significado do vocabulário (RDF ou OWL) predefinido.

3.8. Considerações finais

Ao concluir este capítulo verificou-se que os estudos desenvolvidos com relação às ontologias vêem demonstrando um potencial promissor no desenvolvimento da Web, pois além de seus modelos de estruturação de conceitos e relações, possui um conjunto de axiomas formais que, em forma de uma teoria lógica, permitem a representação formal em nível de significação (nível ontológico). Como conseqüência do formalismo empregado, muitos são os benefícios alcançados, como por exemplo, a verificação/validação automática do modelo de conhecimento construído, a interpretação não ambígua das definições de conceitos e relações e a possibilidade de geração sistemática (e idealmente automática) de infra-estruturas de domínio.

Essa base de definições e relacionamento permite o desenvolvimento de uma Web muito mais significativa, como será apresentado no próximo capítulo.

4 DESENVOLVENDO A SEMÂNTICA NA WEB

Originalmente, o computador era visto somente como hardware. Na década de 80, transformou-se em um sistema capaz de simular jogos, processar textos e elaborar apresentações. Hoje em dia, tornou-se um portal para uma rede de troca de informações e transações comerciais. Como conseqüência, as tecnologias que dão acesso a essas informações textuais, não estruturadas e heterogêneas se tornaram tão essenciais quanto às linguagens de programação nas décadas de 60 e 70. A Internet, mas especificamente a tecnologia Web, deu início a estas mudanças e acarretou uma série de transformações de caráter tecnológico, social e econômico. A Web passou a propiciar uma nova plataforma para o desenvolvimento de aplicações com acesso distribuído. Antes de seu surgimento, os principais serviços utilizados na Internet eram as transferências de arquivos, o correio eletrônico e a emulação de terminal, e restritos aos meios acadêmicos e militares. O uso generalizado da Internet só veio a acontecer, em 1992, com o surgimento da Web, que organizou as informações na Internet por meio de hipertextos e, em um segundo momento, tornou a interação do usuário com a rede mundial mais amigável.

Segundo (DACONTA, 2003), no início, a Web era um projeto desenvolvido, a partir de março de 1989, por Tim Berners-Lee[2] no *European Organization for Nuclear Research* (CERN[3]), para acessar informações estanques espalhadas pelos diversos laboratórios na Europa, tendo evoluído para um serviço usado globalmente. O que era um sistema baseado em buscas por hipertexto teve

[2] http://www.w3.org/People/Berners-Lee/
[3] http://user.web.cern.ch/user/cern.html

seu crescimento viabilizado pelo traço cooperativo da Internet, ou seja, pela colaboração mútua entre os componentes da rede. A aparente simplicidade da Web está criando obstáculos para seu próprio desenvolvimento, já que a tecnologia utilizada atualmente limita a manipulação da informação.

O primeiro objetivo do projeto da Web era criar um ambiente em que pudéssemos trabalhar melhor em grupo tanto no trabalho quanto em casa. A idéia era que criando uma rede de hipertextos, os grupos de usuários seriam forçados a utilizar um vocabulário comum entre eles para que não ocorressem mal entendidos e, em algum momento, teriam um modelo na Web dos planos e idéias em discussão no grupo. O precursor da Web foi um programa para uso próprio, chamado "*Enquire*", desenvolvido por Tim Berners-Lee, em 1980, quando trabalhava no CERN. Este programa tinha o propósito de manter registros da complexa rede de relacionamentos entre pessoas, programas, máquinas e idéias espalhadas pelos diversos laboratórios na Europa. Mais tarde, em 1989, ele viria a apresentar uma proposta para a Web que, na verdade, era uma extensão deste programa pessoal. (DACONTA, 2003)

Para o seu pleno funcionamento, a Web não tem que ser apenas fácil de se navegar, mas também auto-explicativa. Qualquer informação disponível na Web pode ser facilmente assimilada e qualquer informação que esteja faltando pode ser facilmente adicionada. A Web deve ser um meio de comunicação entre as pessoas, servindo com um meio de compartilhamento do conhecimento. Isso requer que computadores, redes, sistemas operacionais e programas sejam transparentes aos usuários, disponibilizando somente uma interface intuitiva e o mais direta possível com a informação.

(DACONTA, 2003) comenta que o segundo objetivo da Web é baseado na premissa que se há informação disponível na Web então é possível estruturarmos esta informação, criando um mapa de relacionamentos e dependências. Isso possibilitaria o acesso dos programas a estas informações e permitiria que eles nos ajudassem em sua análise e gerenciamento. A estruturação do conteúdo semântico da informação das páginas web criaria um ambiente onde agentes de softwares executam tarefas solicitadas pelos usuários, deixando a cargo dos computadores qualquer tarefa que possa ser reduzida a um processo racional.

Para atingir os objetivos de criação de uma Web de acesso universal e que contenha informações estruturadas de maneira a serem utilizadas pelas máquinas na automação de tarefas e informações confiáveis em que possam ser identificados os autores e responsáveis por suas publicações, o W3C tem como foco os seguintes objetivos[4]:

- Acesso universal;
- Web Semântica;
- Web confiável;
- Interoperabilidade;
- Evolução;
- Descentralização;
- Ambiente interativo.

[4] Detalhes dos objetivos e princípios operativos do W3C em http://www.w3.org/Consortium/Points/

4.1 Web Semântica

Existem duas perspectivas distintas para o futuro da Web. A primeira perspectiva se preocupa em intensificar a colaboração na Internet, enquanto a segunda perspectiva está voltada para uma Web mais compreensível e, portanto mais eficiente, para máquinas Figura 4.1.

Mesmo a proposta original da Internet envolvia mais do que a recuperação de páginas HTML a partir de servidores web. A Figura 4.1 ilustra as relações

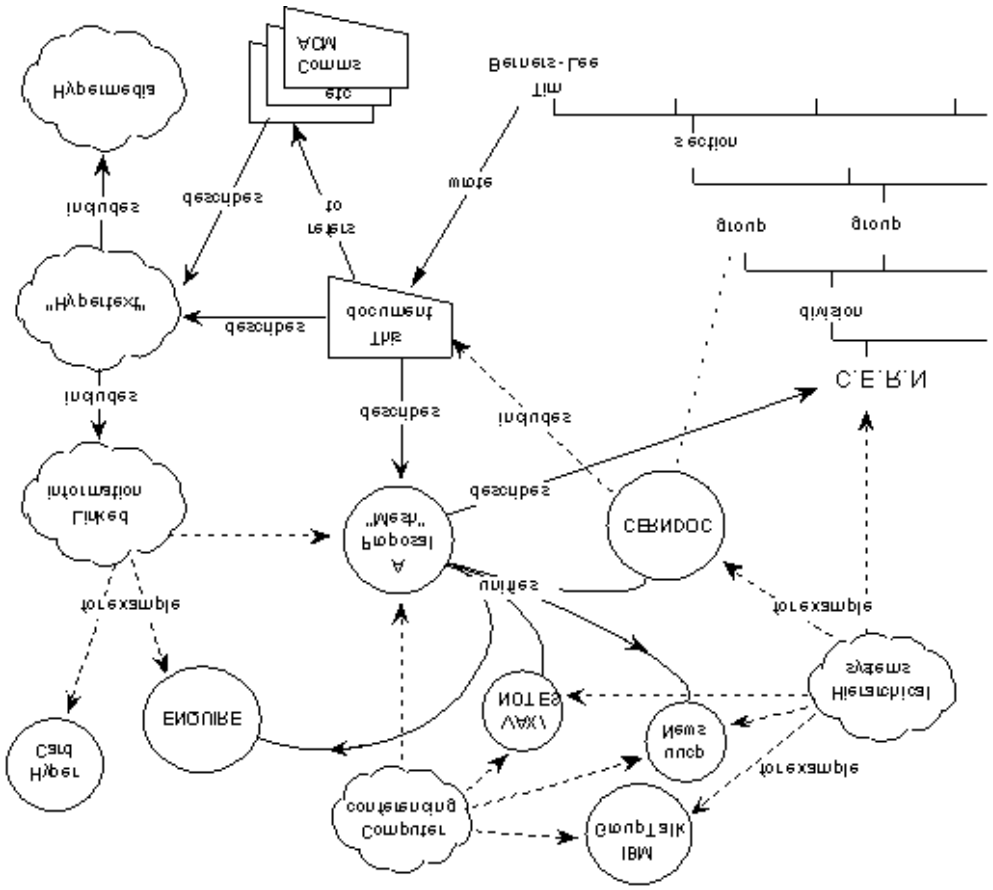

Figura 4.1 Proposta original da Web

Copyright © Tim Berners-Lee

entre alguns itens de informação, tais como "*includes*", "*describes*" e "*wrote*". Infelizmente esses relacionamentos entre os recursos não são atualmente capturados na Web. A tecnologia para capturar esses relacionamentos chama-se RDF. O ponto chave para compreender a Figura 4.1 e que a visão original englobava metadados[5] adicionais acima e além do que atualmente está na Internet.

(DACONTA, 2003) comenta que a questão resume-se a como uma criar uma rede de dados que as máquinas possam processar. Para atingir este fim, o primeiro passo é uma mudança de paradigma na maneira como as pessoas pensam sobre dados. Historicamente, os dados eram isolados em aplicações proprietárias, sendo paradoxalmente considerados recursos secundários para o processamento de dados. Esta visão incorreta popularizou a expressão *Garbage in, Garbage out* (GIGO). O GIGO basicamente revela a falha no argumento original estabelecendo a dependência entre processamento e dados. Em outras palavras, o software útil era completamente dependente de bons dados. Os profissionais de informática começaram a perceber que os dados eram importantes, e que deveriam ser verificados e protegidos. As linguagens de programação começaram a implementar os conceitos de orientação a objetos[6] que internamente transformam dados em estruturas de maior significado.

Entretanto, essa abordagem que enfatiza a importância do dado foi mantida interna às aplicações de tal forma que os fornecedores poderiam manter os dados proprietários por razões competitivas. Com o advento da Internet, da

[5] Metadados adicionais são necessários para que as máquinas processem informações na Web.
[6] Detalhes de orientação a objetos aplicados a Web em http://www.w3.org/OOP/

XML e, mais recentemente da web semântica o valor até então agregado às aplicações deslocou-se para os dados. Este fato define essencialmente o que é a web semântica. O caminho para tornar os dados mais compreensíveis as máquinas é torná-los mais inteligentes. Todas as tecnologias desse trabalho são os fundamentos de uma abordagem sistemática para criar dados inteligentes. A Figura 4.2 mostra os quatro estágios da evolução dos dados, que se tornam continuamente mais inteligentes. Esses estágios evoluem de dados minimamente inteligentes a dados embutidos de informações semânticas suficientes para que as máquinas possam fazer inferências. (DACONTA, 2003)

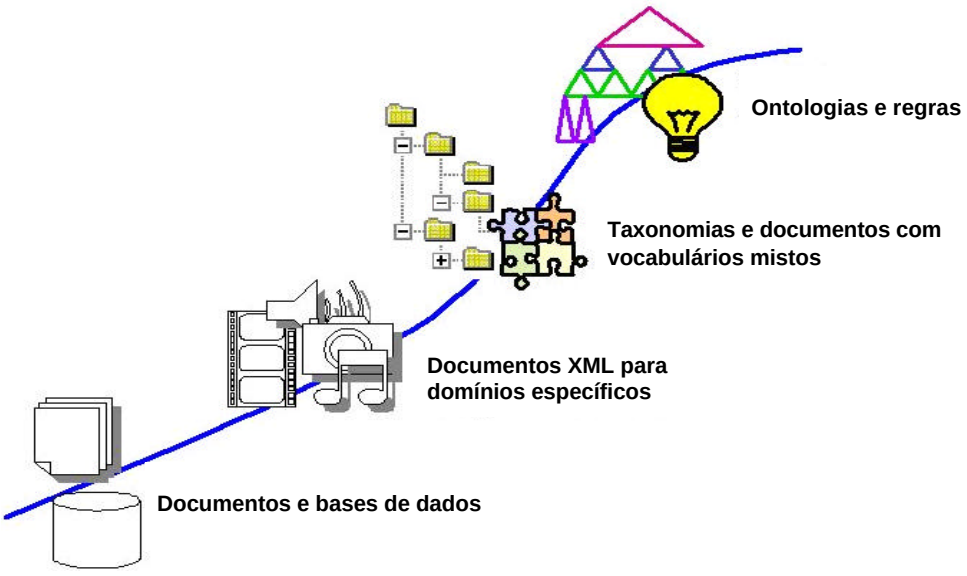

Figura 4.2 Evolução dos dados

O detalhamento dos quatros estágios é apresentado a seguir:

- **Texto e banco de dados (pré-XML):** a maioria dos dados eram propriedades de uma aplicação. Portanto, a inteligência estava na aplicação e não nos dados.

- **Documentos XML para domínios específicos:** os dados se tornam independentes da aplicação dentro de um domínio específico. Os dados agora são inteligentes o suficiente para se moverem entre aplicações de um domínio. Um exemplo disso poderiam ser os padrões XML da industria química, automobilística, aeroespacial etc.

- **Taxionomias e documentos com vocabulários mistos:** os dados podem ser compostos de múltiplos domínios e podem ser classificados dentro de uma taxionomia hierárquica. De fato, a classificação pode ser utilizada para descoberta de dados. Relacionamentos simples entre categorias na taxonomia podem ser usados para relacionar e combinar dados. Portanto, os dados são inteligentes o suficiente para serem facilmente descobertos e sensivelmente combinados com outros dados.

- **Ontologias e regras:** novos dados podem ser inferidos a partir dos existentes seguindo regras lógicas simples. Essencialmente, os dados são suficientemente inteligentes para serem descritos com relacionamentos concretos, e formalizam sofisticados onde cálculos lógicos podem ser realizados com esta “álgebra semântica”. Isso permite a combinação e recombinação de dados até um nível atômico e uma análise mais refinada dos dados. Portanto, neste estágio, os dados não existem como uma bolha, mas sim como parte de um microcosmo sofisticado. Um exemplo dessa sofisticação dos dados é a tradução automática de documentos de um domínio para documentos equivalentes em outros domínios.

Neste ponto, é possível estabelecer uma nova definição para web semântica: redes de dados inteligentes que são processáveis pelas máquinas. De uma maneira mais detalhada, pode-se definir dados inteligentes como dados que: não dependem de aplicação, classificáveis, de fácil composição e que fazem parte de um ecossistema de informação maior (Ontologias).

4.2 Motivação para web semântica?

Segundo (DACONTA, 2003) a web semântica não envolve apenas a WWW. Ela representa um conjunto de tecnologias necessárias às atividades desenvolvidas em redes coorporativas, *intranets*. Isto é análogo aos w*eb services*, que representam serviços não somente através da Internet, mas também nas *intranets*. Nesse sentido, a web semântica propõe solucionar vários problemas chaves das atuais arquiteturas de tecnologia da informação como:

- **Sobrecarga de informação:** as ferramentas de busca enfrentam a dificuldade de executar pesquisas entre documentos que não estão diferenciados em termos de assunto, qualidade e relevância. A tecnologia atual não é capaz de diferenciar uma informação comercial de um educacional, ou informação entre idiomas, culturas e mídia. É necessário haver informações de qualificação da própria informação para que seja possível classificá-las e tornar os processos de recuperação de informações mais eficazes.

- **Integração de informações:** a variedade de fontes de informação distintas com diferenças sintáticas, semânticas e estruturais é muito

grande, tornando o compartilhamento, integração e resolução de conflitos entre informações um problema difícil de ser solucionado. Outra questão a ser tratada seria a criação ou remoção de fontes de informação, o que teria que ser realizada com extrema cautela de forma a não causar grandes impactos ao ambiente integrado. Deve-se considerar que as fontes de informação podem ter capacidades computacionais diferentes, podendo variar desde sistemas de banco de dados a arquivos. As informações podem variar de não estruturadas, como imagens e vídeos, as semi-estruturadas, como arquivos de e-mail e páginas Web. A heterogeneidade estrutural e semântica da informação na Web, atualmente, é imensa e a maioria das propostas de integração ainda adota soluções com alto índice de centralização, tornando seu uso na Web inviável.

- **Conteúdo não estruturado:** um dos motivos do grande sucesso da Web é sua liberdade de publicação de informação. Essa liberdade proporcionou uma enorme quantidade de documentos e recursos de todo tipo disseminado na Web, tais como: bancos de dados, artigos, programas, arquivos, etc. Por serem criados de forma autônoma, sem preocupação com regras de estruturação, catalogação e descrições de suas propriedades, essas informações são difíceis de serem abrangidas pelos mecanismos de pesquisa, ocasionando demora e ineficácia na localização de informações. Alguns problemas enfrentados pelos mecanismos de busca e recuperação de informações são: demora na localização de informações; informações não localizadas devido às mudanças de URL; e, recuperação de informações fora do contexto

solicitado pelo usuário devido a problemas de semântica e ambigüidade. A efetividade dos mecanismos de busca depende principalmente da maneira pela qual as informações foram estruturadas e catalogadas na Web. Documentos podem ser estruturados e organizados de várias formas diferentes na Web e as ferramentas de busca têm que utilizar mecanismos de recuperação adequados para cada tipo de organização.

4.3 Arquitetura da web semântica

Na proposta de desenvolvimento da web semântica (BERNERS – LEE et al., 2001a) sugere uma arquitetura de três camadas conforme Figura 4.3:

- **Esquema:** que estrutura os dados e define seu significado;
- **Ontologia:** que define as relações entre os dados;
- **Lógica:** define mecanismos para fazer inferências sobre os dados.

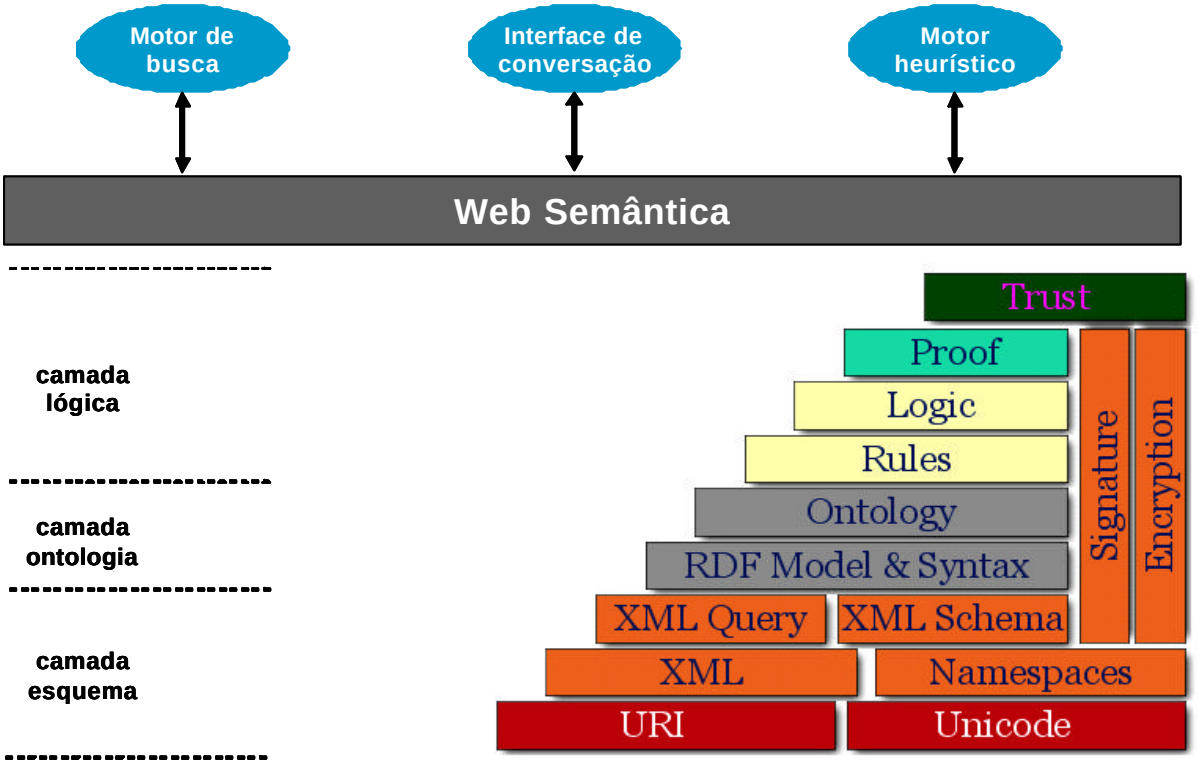

Figura 4.3 Arquitetura da Web Semântica

4.3.1 Camada esquema

A representação do conhecimento é o primeiro passo em direção a web semântica. É necessário que os dados sejam estruturados e que sejam atribuídos significados, para que seja possível elaborar um raciocínio lógico.

Para que haja a representação do conhecimento é necessária à satisfação de três condições:

- **Interoperabilidade estrutural:** provê a representação para modelos de dados distintos, permitindo especificar tipos e possíveis valores para cada forma de representação;
- **Interoperabilidade sintática:** provê regras precisas para promover o intercâmbio dos dados na Web;

- **Interoperabilidade semântica:** possibilita a compreensão dos dados e suas associações com outros dados.

A camada esquema irá controlar os dados nos documentos, como uma gramática de definição, em que os mesmos estarão estruturados e com significados bem definidos. As tecnologias base dessa camada são descritas a seguir:

- ***Uniform Resource Identifier*** **(URI[7]):** é simplesmente um identificador da Web. O URI é utilizado para identificar itens na Web. Uma *Uniform Resource Locator* (URL) é um tipo de URI que fornece um caminho para obter mais informação sobre determinado recurso. O URI é uma tecnologia ideal para ser utilizada *pelo* RDF[8].

- **XML:** as páginas da Web na sua maioria são baseadas em HTML[9]. Em uma linguagem pré-definida como o HTML não se tem à possibilidade de estender o conjunto de *tags* original, isto é, caracterizar os documentos; com essa facilidade, as aplicações específicas poderiam associar significados aos dados e aos campos dos documentos o que viabilizaria seu processamento automático, indo muito além do que simplesmente gerar uma visualização. A solução para a extensão do conjunto de *tags* foi produzir o XML[10] (WEIBEL, 1997). A definição da linguagem XML consiste em um padrão utilizado para marcação de documentos que contém informações estruturadas, ou seja, documentos que contém uma estrutura clara e precisa da informação que é armazenada em seu

[7] http://www.w3.org/Addressing/
[8] http://www.w3.org/RDF/
[9] Linguagem descendente do SGML que é uma meta-linguagem padrão utilizado com sucesso há mais de uma década para geração de linguagens de representação de documentos.
[10] http://www.w3.org/XML/

conteúdo (WEIBEL, 1997). A estrutura da XML pode ser exemplificada na Figura 4.4.

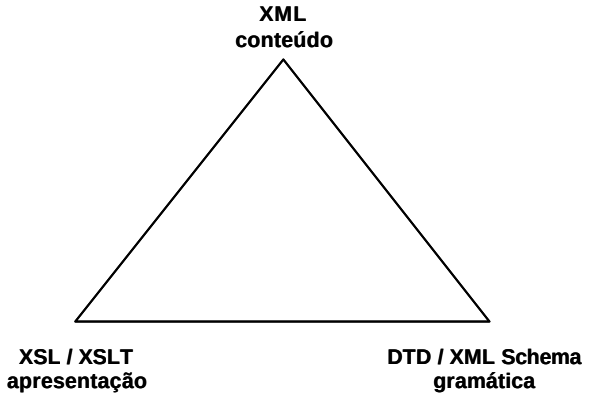

Figura 4.4 Estrutura da linguagem XML

- ***Data Type Definition* (DTD) / XML *Schema*[11]:** gramáticas que definem as regras para a verificação do documento na sua forma sintática e também definem os elementos que constituem a estrutura do documento, assim como seus relacionamentos;
- ***eXtensible Style Language* (XSL[12]) / XSL *Transformations* (XSLT[9]):** linguagens da família XML responsáveis pela apresentação dos documentos XML;
- **XML:** provê o formato dos dados para documentos estruturados sem especificar um vocabulário real, o que irá identificar esse vocabulário será o DTD / XML *Schema*.

A linguagem XML consegue atender duas das três condições para que haja a representação de conhecimento:

[11] http://www.w3.org/XML/Schema
[12] http://www.w3.org/Style/XSL/

- **interoperabilidade estrutural:** realizada através do DTD / XML *Schema*, que provê uma representação dos dados, especificando tipos e possíveis valores.

- **interoperabilidade sintática:** realizada através do DTD / XML *Schema*, que provê regras precisas, utilizando uma gramática, que estabelece a sintaxe.

4.3.2 Camada Ontologia

É na camada de ontologia que garantimos as definições únicas dos conceitos, mesmos os conceitos expressos de formas diferentes e em linguagens diferentes. Essa garantia é atingida através da utilização das triplas <objeto, atributo, valor> nas linguagens. A seguir será detalhada a linguagem RDF por ser a linguagem base das representações ontológicas.

RDF: é uma recomendação do W3C para definir uma padronização da representação e do uso de metadados na Web. O propósito da W3C era fornecer meios para possibilitar a interoperabilidade entre aplicações na Web, permitindo que as informações fossem entendidas diretamente pelas máquinas. (SAUNDERS, 1995).

Segundo (W3C), o RDF define um mecanismo para descrever recursos na Web de uma forma neutra, sem descrever uma área de aplicação específica ou domínio de conhecimento. Em princípio, o RDF não define a semântica de nenhum domínio.

A sintaxe do RDF utiliza a linguagem XML para expressar o significado da informação. Conseqüentemente, a XML e o RDF se tornam complementares.

Enquanto que a XML define uma estrutura, o RDF permite expressar o significado associado aos dados. A flexibilidade da XML permite expressar a semântica através da representação padronizada das triplas <objeto, atributo, valor>, escritas sob a estrutura sintática do XML.

O modelo de dados RDF é definido como:

- Recursos
- Literais
- Propriedades
- Sentenças

As sentenças são formadas por <objeto, atributo, valor> onde:

- **Objeto:** é um recurso;
- **Atributo:** é uma propriedade;
- **Valor:** é um recurso ou um literal.

O relacionamento entre um recurso e um literal é chamado de sentença. A sentença relaciona um atributo a um valor através de um objeto. Exemplo:

- **Sentença:** Bill Gates é dono da http://www.microsoft.com.
    - **Atributo:** dono
    - **Objeto:** http://www.microsoft.com
    - **Valor:** Bill Gates

Os recursos que representam os objetos das sentenças devem utilizar identificadores no padrão URI, pois representam um endereço único para cada recurso na Web.

As sentenças RDF podem ser representadas de três formas distintas:

– **Grafo:** A representação gráfica é a mais fácil de ser compreendida como mostra a Figura 4.5.

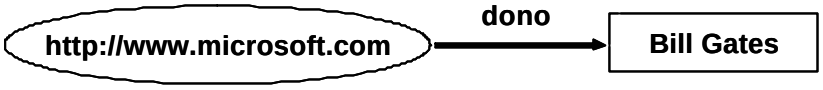

Figura 4.5 Sentença RDF representada na forma gráfica

– **XML:** As sentenças RDF escritas em XML habilitam o intercâmbio entre máquinas, sem interferência humana, através de várias aplicações e

```
<rfd:RDF xmlns:dc= "http://purl.org/metadata/dublin_core#">
    <rdf:Description about= "http://www.microsoft.com">
        <dc:Creator>Bill Gates</dc:Creator>
    </rdf:Description>
</rfd:RDF>
```

Figura 4.6 Sentença RDF escrita em XML

serviços conforme ilustra a Figura 4.6.

– **Triplas:** As triplas são acessíveis às aplicações que irão utilizá-las como entradas para suas operações conforme demonstrado na Figura 4.7.

Recurso	Propriedade	Literal
http://www.microsoft.com	dono	Bill Gates

Figura 4.7 Sentença RDF na forma de triplas

A especificação da RDF não é suficiente para descrever metadados de uma forma associada, representando semanticamente um dado domínio de conhecimento. Quando citamos a representação semântica, queremos nos referenciar principalmente à definição de ontologias e conseqüentemente de todas as características das informações por elas requeridas como definição de

significado, definição de possíveis relacionamentos, propriedades, herança de propriedades e restrições de valores.

A W3C definiu o conceito de esquema RDF para expressar a estrutura de metadados de documentos na Web. Para tal é utilizado o modelo RDF associado a regras, formando um vocabulário do esquema, que define a estrutura dos dados, de forma similar a um esquema tradicional de banco de dados. Logo, os dados definidos de acordo com o esquema estarão sempre em conformidade. A Figura 4.7 apresenta um pequeno exemplo da estrutura do RDF *Schema*.

```
<rdf:RDF xml:lang="en"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
    <rdf:Description ID="PosGraduacao">
        <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
        <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    </rdf:Description>
    <rdf:Description ID="Especializacao">
        <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
        <rdfs:subClassOf rdf:resource="#PosGraduacao"/>
    </rdf:Description>
    <rdf:Description ID="Mestrado">
        <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
        <rdfs:subClassOf rdf:resource="#PosGraduacao"/>
    </rdf:Description>
    <rdf:Description ID="Doutorado">
        <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
        <rdfs:subClassOf rdf:resource="#PosGraduacao "/>
    </rdf:Description>
</rdf:RDF>
```

Figura 4.8 Exemplo da sintaxe do RDF *Schema*

A linguagem RDF satisfaz as três condições para que haja a representação de conhecimento:

- **interoperabilidade sintática:** através do RDF *Schema*, ou seja, provê regras precisas, utilizando uma gramática, que estabelece uma sintaxe;

- **interoperabilidade estrutural:** provendo uma representação aos dados e especificando tipos e possíveis valores para cada forma de representação, através do mecanismo de <objeto, atributo, valor>;
- **interoperabilidade semântica:** através do significado que é atribuído aos dados, através das triplas de <objeto, atributo, valor>.

4.3.3 Camada Lógica

A camada lógica é composta por um conjunto de regras de inferência que os agentes poderão utilizar para relacionar e processar a informação. As regras de inferência fornecem aos agentes o poder de analisar os termos e os seus significados, que foram definidos na camada esquema e de analisar os relacionamentos entre os conceitos segundo sua definição na camada ontologia. “Um agente é um sistema computacional que está situado em um ambiente e que é capaz de atuar de forma autônoma neste ambiente com intenção de atingir os objetivos de seu desenvolvedor“ (WOODRIDGE, 1999).

Os agentes possuem algumas características que os distinguem: têm autonomia (operam, tomam decisão sem intervenção externa), têm reatividade (são ativos, percebem o seu ambiente e agem), têm comportamento colaborativo, possuem objetivos (metas), são flexíveis, sociáveis e têm capacidade de aprender. Os agentes serão capazes de “compreender” o significado entre objetos, com base em ontologias e de raciocinar sobre eles utilizando regras de inferência definidas na camada lógica. Alem disso, deverão ser capazes de trocar dados entre si.

A Web Semântica terá vários agentes interagindo, cooperando e formando cadeias de valor que facilitem a comunicação e a ação humana. Os agentes poderão trocar ontologias e adquirirem novas capacidades, quando descobrirem novas ontologias (BERNERS – LEE et al., 2001b).

4.4 O papel dos *web services* na web semântica

O uso de w*eb services* está crescendo rapidamente por possibilitar de forma transparente a interoperabilidade e comunicação entre aplicações. Estes serviços provêem uma arquitetura para implementar a comunicação entre aplicações de diferentes tipos e plataformas.

Esta arquitetura focaliza-se em como os serviços são descritos e organizados dinamicamente, propiciando descoberta e uso automático Figura 4.8. Excluem-se aqui sistemas cujas interações são acopladas manualmente de forma rígida, como é o caso nas aplicações de *Electronic Data Interchange* (EDI).

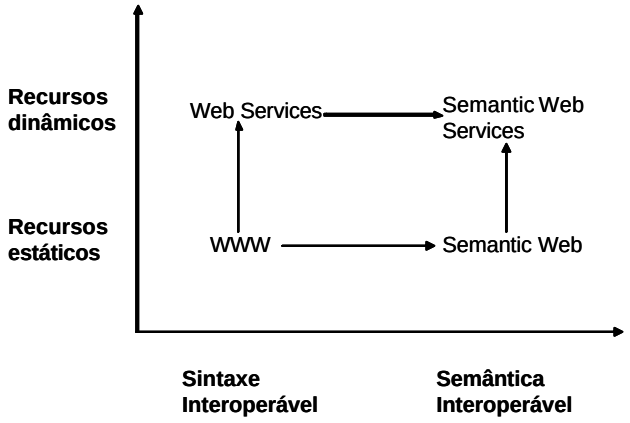

Figura 4.9 Evolução da tecnologia de w*eb services*

Segundo (W3C, 2002), os w*eb services* são *softwares* identificados por uma URL, sendo que a maioria das implementações é baseada no protocolo HTTP, do qual as *interfaces*, WSDL, podem ser descritas e descobertas por

instrumentos baseados em XML e ainda suportar interações diretas com outros *softwares* utilizando mensagens em XML, SOAP, transportadas sobre protocolos de Internet.

4.4.1 Protocolos utilizados pelos Web Services

Segundo (WEB SERVICES, 2002), os aplicativos acessam *os web services* utilizando os protocolos padrões da Web, sem a necessidade de se preocuparem como cada w*eb service* foi implementado. Todos esses padrões são definidos pelo W3C, tornando a sua implementação viável em qualquer plataforma.

Para implementar w*eb services*, a W3C define quatro padrões, Figura 4.9, primários que são (WEB SERVICES, 2002):

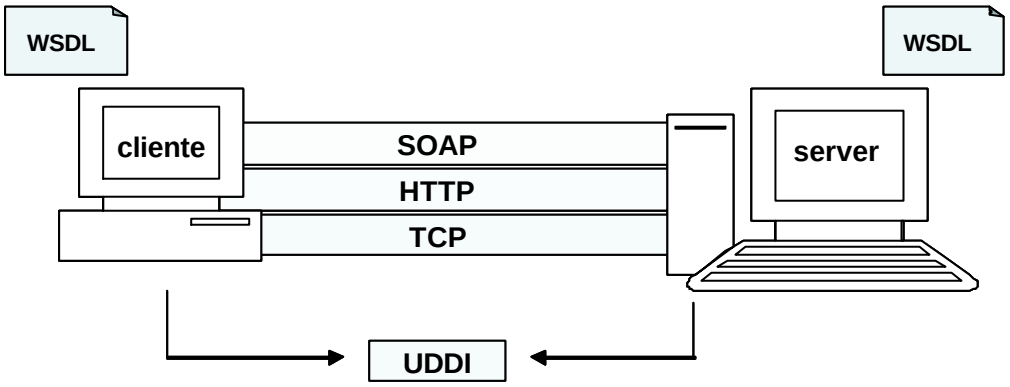

Figura 4.10 Esquema dos w*eb services* com os protocolos

1. **XML:**. a XML é a linguagem dos *Web Services*. Ela é equivalente ao *American Standard Code for Information Interchange* (ASCII) para a Internet, onde dados ou conteúdo que são expressos em XML devem ser entendidos por qualquer aplicativo que entende XML (MARTIN, 2001);

2. **SOAP (*Simplified Object Access Protocol*):** protocolo utilizado para chamar procedimentos remotos pela Internet. É um documento XML que define, entre outras coisas, como uma aplicação pode dizer a um servidor que determinado objeto deve ser carregado, qual método deve ser executado, com quais parâmetros e qual é o valor de retorno;

3. **WSDL (*Web Services Description Language*):** é um documento XML que auxilia na descrição dos serviços. Proporcionam aos clientes de um w*eb services* todos os detalhes necessários para conexão e utilização dos w*eb services*. (Web Services, 2002) e;

4. **UDDI (*Universal Description, Discovery and Integration*):** para transformar a atual Web semi-estática numa Web dinâmica, dotada de serviços colaborativos, é necessário criar um diretório mundial para consultas dos serviços disponíveis, que as empresas divulgarão e procurarão serviços de forma dinâmica. Tal como uma lista telefônica, o UDDI contém páginas brancas e amarelas que permitem procurar os serviços em função do nome da empresa ou do tipo de atividade que exercem. Porém, o UDDI oferece, também, páginas verdes, e indicam como negociar com as empresas que figuram no diretório, designando, por exemplo, os seus procedimentos comerciais ou descrevendo os serviços que disponibilizam. Brancas, amarelas ou verdes, as páginas UDDI possuem um modelo de dados baseado em XML, e o seu acesso exige a utilização do SOAP, tanto para as preencher como para pesquisar os serviços que as contêm (UDDI, 2002).

4.5 Considerações finais

Um dos objetivos originais da Web era a troca de informação entre pessoas, mas de forma de que as máquinas pudessem participar da comunicação, ajudando os usuários. Os computadores na Web, atualmente, têm papel somente no direcionamento e entrega de informações, não tendo acesso ao conteúdo das páginas, porque essa informação está estruturada para utilização pelas pessoas e não por máquinas.

Este capítulo abordou a tecnologia da web semântica, uma proposta de extensão da Web atual baseada no uso de ontologias para descrever relacionamentos entre objetos, formados com informações semânticas, para automatizar o processamento pelas máquinas.

A web semântica não é apenas uma ferramenta para conduzir e auxiliar a execução de tarefas individuais e pesquisas mais eficientes na Web, mas também uma ferramenta para assistir no desenvolvimento do conhecimento.

Devido a sua importância no desenvolvimento da próxima geração da Web, o próximo capítulo irá apresentar uma proposta de arquitetura de integração semântica de bibliotecas digitais utilizando a web semântica como base de desenvolvimento.

5 INTEGRADOR SEMÂNTICO DE BIBLIOTECAS DIGITAIS

A publicação de textos completos na rede, principalmente de Ciência e Tecnologia (C&T), já é uma realidade no Brasil conforme demonstram algumas iniciativas como a SciELO[13], Portal do Conhecimento[14] da USP, Biblioteca Digital LAMBDA[15] da PUC/RIO, Banco de Teses e Dissertação BTD[16] do PPGEP/UFSC, Biblioteca Digital de Teses e Dissertações Brasileira – BDTD[17] do IBICT entre outras. A presença desse novo modelo de biblioteca na sociedade contribuiu para o surgimento do termo biblioteca digital.

(BAX, 1997) traz a seguinte definição de bibliotecas digitais:

> "*as bibliotecas digitais são entidades capazes de vencer as limitações naturais, espaço - temporais, impostas a objetos físicos (livros, estantes, salas, prédios), permitindo novas práticas de trabalho e oportunidades. [...] é uma reunião de um ferramental de computação, estoque e comunicação digitais juntamente com o conteúdo e software necessário para se reproduzir, emular, estender os serviços oferecidos por bibliotecas convencionais baseadas em papel e outros meios de coleção, catalogação, e disseminação da informação. Uma biblioteca digital completa deve ser capaz de oferecer todos os serviços essenciais de uma biblioteca tradicional, assim como explorar as bem conhecidas vantagens do estoque, pesquisa e comunicação digital.*".

[13] http://www.scielo.br
[14] http://www.teses.usp.br
[15] http://www.lambda.maxwell.ele.puc-rio.br
[16] http://www.teses.eps.ufsc.br
[17] http://www.ibict.br/bdtd

(DRANBESTOTT, 1997) comenta que:

> "*...ao se levar em conta outras características e mecanismos do que se denomina biblioteca digital, encontram-se termos complementares, tais como acessibilidade local, nacional, regional, universal, conexão eletrônica, por meio de computadores massivos e roteadores, transparência das informações, independentemente de local ou determinado campus, laboratório de pesquisa, uso de computadores pessoais e portáteis, instituições, firmas comerciais; usuários cadastrados com posse de senhas*".

(LEVACOV, 1997) comenta que:

> *"Para alguns, significa simplesmente a troca de informações por meio da mídia eletrônica e pode abranger uma grande variedade de aplicativos, [...] para outros, significa a possibilidade de [...] criar uma rede mundial que fosse um grande depositário (potencialmente infinito) de todos os documentos da humanidade".*

A *Digital Library Federation (*DLF[18]*)* apresenta sua definição sobre biblioteca digital com base em características e principalmente exigências funcionais para o desenvolvimento da mesma.

- **Organizações que fornecem os recursos:** são organizações que empregam e indicam uma variedade de recursos, especialmente os recursos intelectuais pertinentes à equipe especializada, mas não necessitam ser organizada no modelo das bibliotecas convencionais.

[18] http://www.diglib.org/

Embora os recursos que as bibliotecas digitais requerem sejam similares àquelas dentro das bibliotecas convencionais, eles são, muitas vezes, de tipos diferentes. Por o exemplo, para o armazenamento e a recuperação, as bibliotecas digitais são dependentes quase exclusivamente do computador, das habilidades eletrônicas dos sistemas da rede e da engenharia dos sistemas.

- **Preservam a integridade e asseguram a persistência:** devido à mudança contínua no ambiente torna-se muito difícil assegurar a persistência e a preservação de integridade. Mas a DLF considera estas funções como centrais ao conceito de biblioteca digital. A integridade de objetos digitais é medida em termos de conteúdo, fixação, referência, origem e contexto (LEVY, 1995). A preservação da integridade do objeto, embora necessário, não é uma condição suficiente da persistência. A persistência depende também de outros fatores: vontade organizacional, meio financeiro e a negociação de direitos legais.
- **Coleções de trabalhos digitais:** as distinções entre as bibliotecas geralmente estão no foco do tema da matéria que define as coleções (por exemplo, medicina, arte, ciência, música, e outros), ou nas comunidades interessadas nos materiais coletados (por exemplo, pesquisa, faculdade, público). Para DLF, com o amadurecimento das bibliotecas digitais, o princípio que define as políticas de suas coleções não será o material digital. Outrossim, a definição dos princípios estará, como em outras bibliotecas, na “matéria tema” das matérias e no interesse protetor da comunidade nelas. A pergunta estratégica chave para as bibliotecas digitais que antecipam tal desenvolvimento será

"como integrar coleções de materiais na forma digital com materiais em outras formas".

- **Prontamente e economicamente disponível:** como em outras organizações, as bibliotecas digitais necessitam desenvolver critérios para medir seu desempenho num ambiente em desenvolvimento e altamente competitivo. No mínimo, devem refletir os atributos funcionais de uma biblioteca digital como descritos anteriormente. Uma medida essencial da qualidade do serviço avalia o desempenho nos termos de custo. Embora os custos de serviços de uma biblioteca digital ainda não sejam bem compreendidos, a DLF comenta que as iniciativas bem sucedidas têm uma certa segurança dos fatores críticos de custo e trabalham rapidamente para economizar a influência destes fatores. Uma segunda medida essencial da qualidade de serviço considera a disposição e compreensão de como uma biblioteca digital deixa a informação disponível à comunidade.

- **Uso por uma comunidade ou por um conjunto de comunidades definidas:** as bibliotecas em geral, e as bibliotecas digitais particularmente, são organizações de serviços. As necessidades e os interesses das comunidades que elas servem determinam a trajetória do desenvolvimento das bibliotecas digitais, incluindo o investimento feito no conteúdo e na tecnologia. A maioria das bibliotecas deve ser dedicada ao suporte da educação e da pesquisa, justificando seu investimento em desenvolvimentos digitais como um meio poderoso de realizar os objetivos institucionais das comunidades.

As definições apresentadas em torno das bibliotecas digitais demonstram o seu papel fundamental no compartilhamento do conhecimento, principalmente científico, como apresentado por (MARCONDES, 2002):

> "*a informação de interesse para a pesquisa científica é em grande parte composta pela chamada documentação não convencional, também chamada de "literatura cinzenta", documentos que não são encontrados no circuito editorial convencional, como relatórios de pesquisa, trabalhos apresentados em eventos, teses e dissertações, que noticiam com grande atualidade os resultados de pesquisa.*
>
> *Coletar esta "literatura cinzenta" sempre foi caro e extremamente trabalhoso para os sistemas, devido à dispersão da literatura científica. O surgimento das publicações eletrônicas começa a mudar radicalmente este quadro. Hoje em dia não é mais suficiente para garantir o máximo de visibilidade de seu acervo que bibliotecas digitais simplesmente disponibilizem seus dados na Internet. A quantidade de informações disponível na rede é tão grande que identificar, localizar, descobrir a existência e acessar informações relevantes torna-se um problema crítico, demandando um tempo proibitivo aos usuários*."

Uma visão mais direcionada a realidade de bibliotecas digitais em universidade é apresentada em (PACHECO, 2001a)

> "*O avanço das novas tecnologias na área de bibliotecas é evidente. No entanto, a realização é muito pequena, se comparada ao potencial realizável. Cada universidade ou cada programa de pós-graduação desenvolve, ainda, iniciativas pouco integradas. Provedores de*

informação privados oferecem soluções também pouco integráveis ao restante do ambiente bibliotecário".

A possibilidade de integração de bibliotecas digitais depende da construção de padrões que permitam o intercâmbio de informações e a interoperabilidade de aplicativos. A busca por interoperabilidade entre arquivos abertos, segundo Sena[16], visa a transformar cada um dos arquivos em parte de um arquivo global para a realização de pesquisas on-line."

Um dos apelos no desenvolvimento das bibliotecas digitais é a maior visibilidade dos acervos por ela atendida. Atingir esta visibilidade não significa mais necessariamente que algum usuário buscando informações deva acessar a biblioteca digital ou arquivo eletrônico para ter acesso aos documentos digitais. A possibilidade dos acervos serem consultados simultaneamente sem que um usuário acesse cada *site* individualmente, a chamada interoperabilidade, vem sendo perseguida como o mecanismo viável. Atingir a interoperabilidade entre repositórios de bibliotecas digitais, distintos e heterogêneos possibilitando consultas simultâneas, envolve um aporte intenso em termos de tecnologias, protocolos e padronização.

5.1 Arquitetura de integração semântica de bibliotecas digitais

Na abordagem de integração proposta foi utilizada à visão clássica das camadas de informação (BERGAMASCHI et al., 1999), (CALVANESE et al., 1998) como ilustra a Figura 5.1. A camada semântica é representada por um modelo conceitual que descreve um domínio particular de interesse, domínio

esse que se encontra implícito nas estruturas das fontes a serem integradas. A camada lógica é representada por um modelo lógico que descreve a estrutura

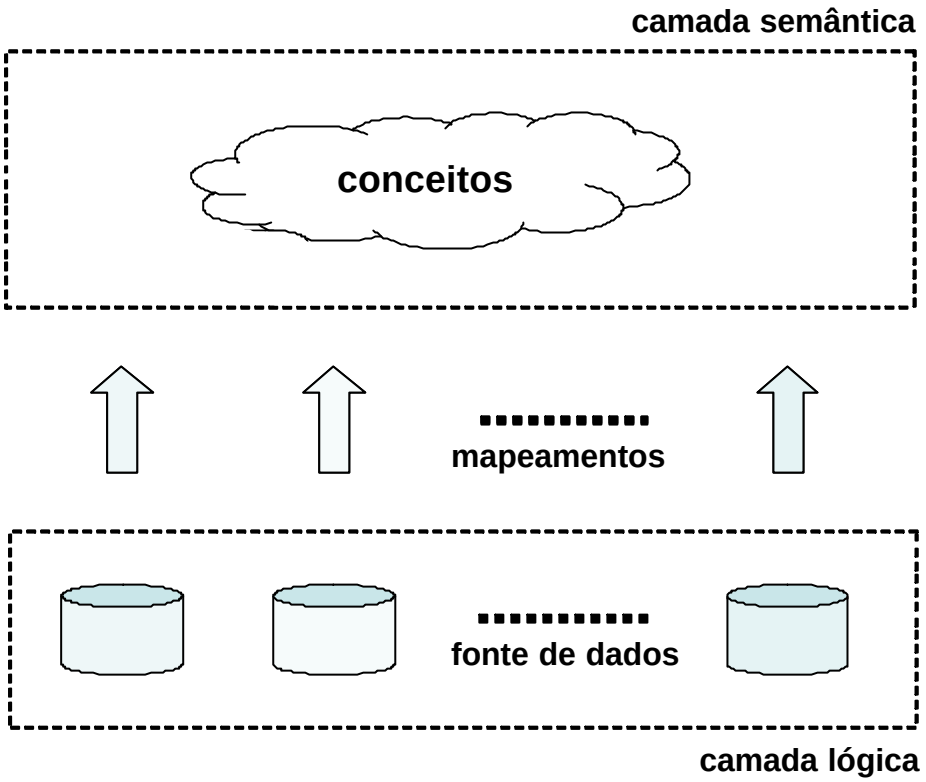

Figura 5.1 Integração estrutural de fonte de dados

das fontes participantes do processo de integração. O mapeamento entre as duas camadas é realizado através do modelo de mapeamento, que contém um conjunto de regras de mapeamento, responsáveis por especificar como elementos do modelo lógico devem ser interpretados no modelo conceitual.

- **Modelo conceitual:** expressa informações semânticas em termos das associações entre conceitos de um domínio particular de interesse, diferentemente das perspectivas ontológicas de representação do conhecimento, que se concentram em expressar a definição de um conceito.

- **Modelo lógico**[19]: tem por objetivo expressar os elementos que compõem uma descrição estrutural das fontes de dados participantes, de forma que cada um destes elementos possa ser mapeado para elementos do modelo conceitual em uma etapa posterior, garantindo uma semântica associada a estes elementos, e conseqüentemente, facilitando o processo de integração destas fontes. Este modelo corresponde a uma simplificação de uma estrutura, incluindo apenas os conceitos de esquema, elemento e domínio de um elemento.

- **Modelo de mapeamento:** tem por objetivo prover informação semântica ao dado estrutural. O modelo compreende dois elementos:

  - um responsável por acomodar um conjunto de restrições, no nível semântico, presentes nos esquemas;

  - um conjunto de regras de mapeamento, as quais são responsáveis por associar o modelo lógico ao modelo conceitual.

O propósito da arquitetura, baseada na visão clássica das camadas de informação, é prover um sistema de integração semântico de dados baseado em *web services* e ontologias através de um esquema comum que sobrepõe a heterogeneidade estrutural dos repositórios de dados estruturados. Este esquema comum é descrito na linguagem OWL e suporta várias ontologias, relacionadas ou não. A OWL é uma linguagem de marcação semântica para recursos da web, desenvolvida como uma extensão da linguagem para ontologias DAML+OIL. Segundo (Smith, 2003), a OWL ultrapassa esta

[19] O termo lógico é aqui utilizado no mesmo sentido empregado em ambientes de bancos de dados, onde denota a descrição de dados em termos das estruturas gerenciadas pelos SGBDs (por exemplo tabelas relacionais), as quais se encontram em um nível mais abstrato com relação à organização física dos dados.

linguagem no que se refere à habilidade para representar na web conteúdo compreensível por máquinas.

É importante salientar que os dados semi-estruturados, páginas HTML, documentos textos e arquivos XML, não foram contemplados na arquitetura. Caso exista necessidade da inclusão de tal fonte de dados, será necessário converter os dados para o formato XML, respeitando a estrutura sintática de acordo com alguma ontologia disponível no repositório de ontologias, Figura 5.2. Após a geração do novo formato os mesmos deverão ser armazenados em algum SGBD.

O diagrama da arquitetura de integração é apresentado na Figura 5.2, sendo seus componentes descritos a seguir.

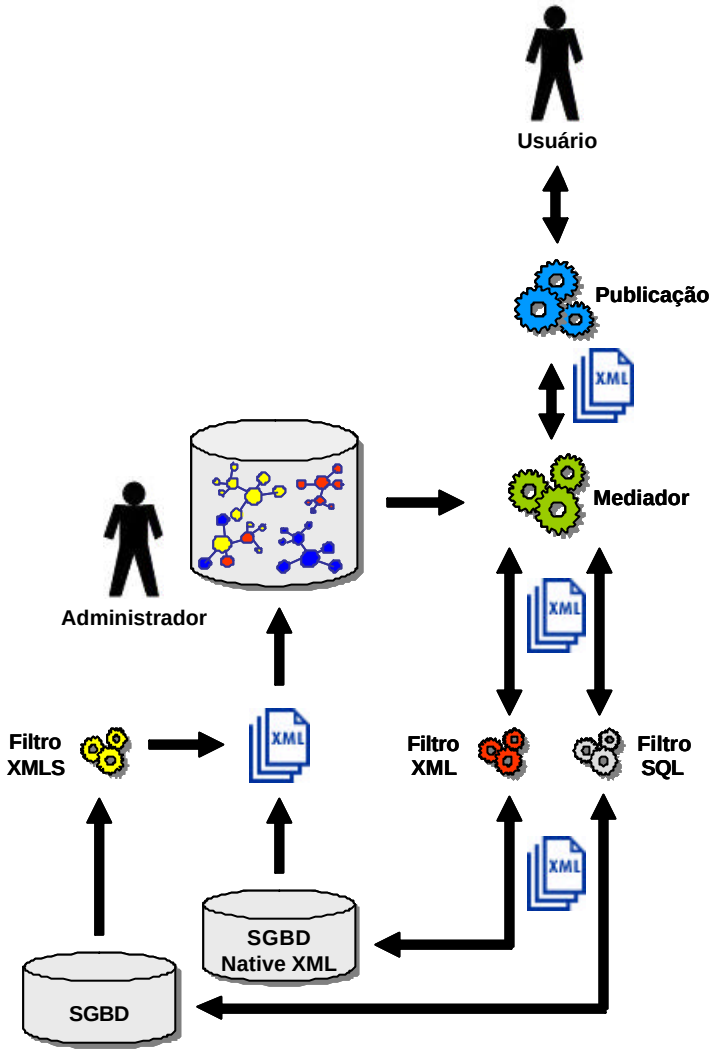

Figura 5.2 Arquitetura proposta para integração semântica

– **Interface usuário:** é responsável pela interação do usuário com a arquitetura proposta. A interface é composta pela interface de publicação, responsável por transformar, converter e publicar os documentos XML nos dispositivos web. Para isso, deverá integrar algumas ferramentas, como processadores e *parsers* de algumas linguagens. É importante lembrar que a separação do conteúdo, lógica e estilo deverá ser respeita, pois essa separação auxilia o processo de apresentação de forma mais eficiente. É através desta interface que os

usuários submetem consultas sem ter conhecimento de como as informações estão de fato representadas.

- **Interface administrador:** recurso responsável pelo ciclo de vida das ontologias. Pode-se dizer que é a parte vital da arquitetura. São tarefas desta interface:

  - **Criação e manutenção do esquema conceitual**: popular e manter o repositório de ontologias com instâncias de conceitos, contendo seus respectivos relacionamentos de dependência e de integração.

  - **Criação e manutenção do catálogo de filtros**: registrar, no repositório de ontologias, a localização dos filtros com a respectiva lista de conceitos envolvidos. Esta informação deve ser obtida no momento em que o esquema de uma fonte de dados é mapeado para repositório de ontologias. Essas informações são de vital importância à arquitetura, pois habilita o mediador a estabelecer comunicação com os filtros de consulta. Nesse estágio é indicada a utilização de arquivos WSDL, pois a funcionalidade é muito semelhante e já é um passo para o caminho da utilização dos *web services* como padrão da arquitetura.

  - **Suporte à tarefa de mapeamento:** disponibiliza a lista de conceitos para que o responsável pelo repositório de dados a ser integrado possa mapear corretamente as fontes de dados que se deseja publicar. Esse processo exige uma atenção especial às

ontologias disponíveis no repositório, já que, o ideal é a reutilização de estruturas pré-definidas e validadas.

As tarefas de criação e manutenção dos esquemas conceituais e de mapeamento são fundamentais para garantir a qualidade do repositório de ontologias, já que são as portas de entradas do repositório.

- **Repositório de ontologias:** é utilizado para armazenar o esquema conceitual que se encontra expresso diretamente em OWL. É importante salientar que a linguagem OWL não é exclusiva na arquitetura. O que é importante é a utilização de linguagens de definição ontológicas baseadas em RDF. A sugestão do uso da linguagem OWL faz-se devido ser esforço mais recente da W3C no que diz respeito a definições de linguagens semânticas.

  A infra-estrutura do repositório de ontologias é baseada na tecnologia de *web services* devido ser a tecnologia que melhor desempenha a funcionalidade de interoperabilidade e por ser completamente descentralizado. Essa realidade de descentralização justifica a utilização da tecnologia de computação em grade no balanceamento da carga de trabalho. O que auxilia de sobre maneira o desempenho do sistema.

  O registro de uma ontologia deve iniciar pelo responsável do repositório de dados a ser integrado. Primeiramente, deverá ser feita uma análise do modelo do repositório ou de uma ontologia em uso com as ontologias disponíveis no repositório de ontologias. Havendo compatibilização ou não com as ontologias do repositório, o administrador do repositório fica responsável pela tarefa de mapeamento ou de criação da ontologia.

Essa responsabilidade é compartilhada com o responsável do repositório a ser integrado.

Após a inserção das ontologias no repositório, as mesmas podem ser dinamicamente atualizadas com novas instâncias de conceitos durante o seu ciclo de vida, vale ressaltar que as ontologias não terminam como o ciclo de vida de software e sim estão em constante evolução. Caso uma ontologia seja descontinuada é porque houve falha na sua constituição.

Além das ontologias, o repositório armazena os catálogos dos filtros. Esses catálogos contêm a localização, a lista de conceitos e o modelo de mapeamento do repositório de dados a ser integrado.

- **Interface mediador:** é responsável por manter uma visão única e integrada ao usuário do sistema. A interface mediador é o centro da arquitetura, pois realiza o processo de orquestração entre usuários, interface de apresentação, repositório de ontologias e filtros da arquitetura. A única exigência dessa interface é a presença da XML como linguagem de comunicação com os componentes da arquitetura.

  Propostas baseadas em mediadores se mostram adequadas no contexto das aplicações em função do fraco acoplamento entre as fontes de dados (visto que não existe um esquema global), respeitando, portanto, a autonomia dos repositórios. O processo de mediação também favorece a construção de aplicações com grau de abstração adequado ao usuário à medida que provê um ponto de acesso único e uniforme as fontes de dados envolvidas.

- **Interface filtro:** apresenta três filtros:

    - **Filtro SQL:** responsável por fazer a comunicação com os SGBDs relacionais através de instruções SQL. Estás instruções são geradas com auxílio da interface mediador que através do acesso ao repositório de ontologias elabora as consultas a partir das definições de mapeamento;

    - **Filtro XML:** responsável por fazer a comunicação com os SGBDs nativos XML através de instruções *XQuery*. A vantagem de utilizar este recurso, *XQuery*, é que a instrução não precisa ser transformada, pois já está no formato XML.

    - **Filtro XMLS:** responsável por fazer o mapeamento da estrutura de um modelo relacional para o padrão XML *Schema* (PADILHA, 2003). A partir do XML Schema o procedimento de conversão ou ajuste para RDF gera pouco trabalho.

Com base na arquitetura ora apresentada foi desenvolvido um protótipo de integração entra as bibliotecas digitais do IBICT e da SciELO. Na próxima seção é apresentado o protótipo.

5.2 Integração das bibliotecas digitais do IBICT da SciELO

A validação da arquitetura de integração semântica de bibliotecas digitais foi feita através do desenvolvimento do protótipo denominado de Base Científica – BASCIN. Esse protótipo fez uso da estrutura da Biblioteca Digital de Teses e Dissertações do IBICT e da Biblioteca Digital de Artigos da SciELO.

Segundo (Marcondes, 2001), o IBICT, através do projeto da Biblioteca Digital Brasileira em C&T - BDB, passou a fomentar o desenvolvimento de recursos informacionais brasileiros de interesse para C&T em texto completo, como teses e dissertações, artigos de periódicos, trabalhos em congressos, arquivos eletrônicos de preprints, integrando e provendo interoperabilidade entre estes recursos através do acesso unificado aos mesmos, via uma única interface web.

A *Scientific Electronic Library Online* - SciELO é uma biblioteca eletrônica que abrange uma coleção selecionada de periódicos científicos brasileiros. O seu objetivo é o desenvolvimento de uma metodologia comum para a preparação, armazenamento, disseminação e avaliação da produção científica em formato eletrônico.

Para o desenvolvimento da infra-estrutura foram seguidas as seguintes etapas:

1. Aquisição dos padrões de metadados das duas bibliotecas digitais.
   - MTD-BR em http://www.ibict.br/schema
   - Artigo DTD-SciELO v3.1 em http://www.scielo.org/dtd/004_pt.htm
2. Mapeamento dos elementos e definições dos dois modelos na linguagem OWL. Para o desenvolvimento do protótipo foi considerado um conjunto mínimo de elementos dos dois padrões, conforme demonstrado na Figura 5.3.

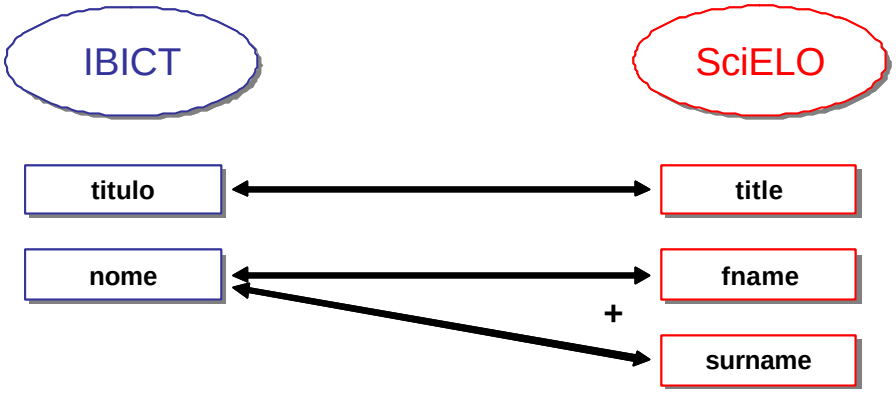

Figura 5.5 Mapeamento dos elementos dos modelos

3. Para codificação foi utilizada a ferramenta OilEd 3.5 que trabalha com a linguagem OIL. A própria ferramenta disponibiliza vários conversores inclusive para linguagem OWL, Figura 5.4. O resultado da codificação em OWL dos padrões pode ser visualizado nas Figuras 5.5 e 5.6.

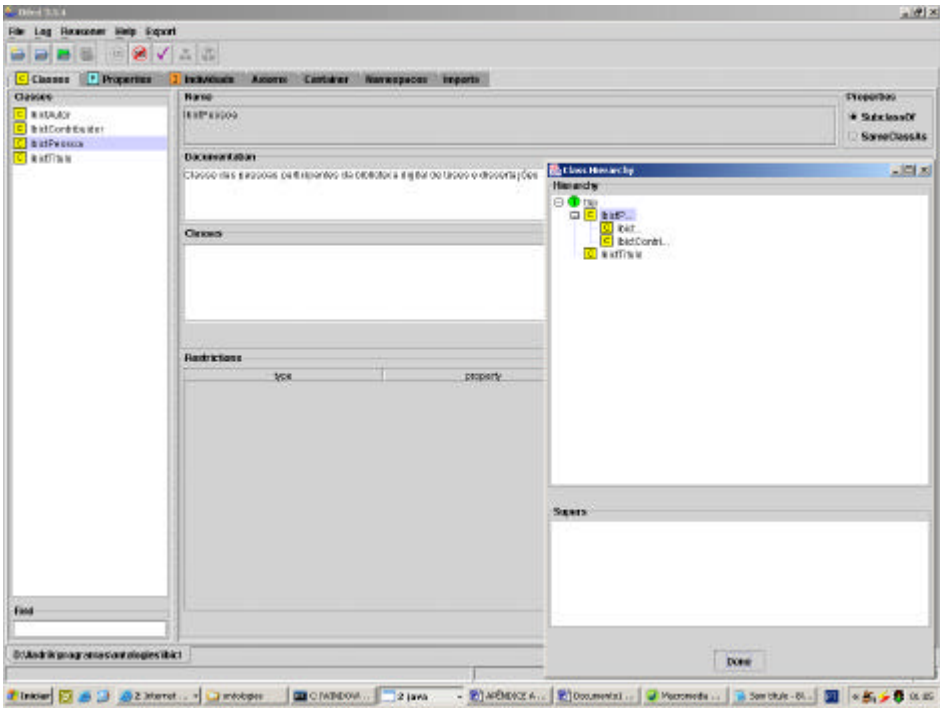

Figura 5.4 Desenvolvimento da especificação OIL do modelo do IBICT

A integração semântica é feita relacionando os modelos OWL da SciELO, Figura 5.5, e do IBICT, Figura 5.6. Esse relacionamento gera um novo arquivo OWL como demonstrado na Figura 5.7

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<owls:Ontology
    xmlns:owls="http://www.w3.org/2003/OWL-XMLSchema"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/OWL-XMLSchema http://wonderweb.semanticweb.org/owl/2003/owl1-dl.xsd">
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloFname"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloTitle"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloSurname"/>
</owls:Ontology>
```

Figura 5.5 Listagem OWL da SciELO

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<owls:Ontology
    xmlns:owls="http://www.w3.org/2003/OWL-XMLSchema"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/OWL-XMLSchema http://wonderweb.semanticweb.org/owl/2003/owl1-dl.xsd">
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictTitulo">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
    </owls:Class>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictAutor">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    </owls:Class>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictContribuidor">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    </owls:Class>
    <owls:ObjectProperty owls:name="file:/D:/ontologies/ibict.daml#e_autor">
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
        </owls:domain>
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
        </owls:domain>
    </owls:ObjectProperty>
    <owls:ObjectProperty owls:name="file:/D:/ontologies/ibict.daml#ibictCitacao"/>
    <owls:ObjectProperty owls:name="file:/D:/ontologies/ibict.daml#e_autor">
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
        </owls:domain>
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
        </owls:domain>
    </owls:ObjectProperty>
</owls:Ontology>
```

Figura 5.6 Listagem OWL do IBICT

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<owls:Ontology
    xmlns:owls="http://www.w3.org/2003/OWL-XMLSchema"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/OWL-XMLSchema http://wonderweb.semanticweb.org/owl/2003/owl1-dl.xsd">
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictTitulo">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
    </owls:Class>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloFname"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloTitle"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictAutor">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    </owls:Class>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/ibict.daml#IbictContribuidor">
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
    </owls:Class>
    <owls:Class owls:complete="false" owls:name="file:/D:/ontologies/scielo.daml#ScieloSurname"/>
    <owls:ObjectProperty owls:name="file:/D:/ontologies/ibict.daml#e_autor">
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictPessoa"/>
        </owls:domain>
        <owls:domain>
            <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictTrabalho"/>
        </owls:domain>
    </owls:ObjectProperty>
    <owls:ObjectProperty owls:name="file:/D:/ontologies/ibict.daml#ibictCitacao"/>
    <owls:EquivalentClasses>
        <owls:Class owls:name="file:/D:/ontologies/ibict.daml#IbictAutor"/>
        <owls:Class owls:name="file:/D:/ontologies/scielo.daml#ScieloSurname"/>
        <owls:Class owls:name="file:/D:/ontologies/scielo.daml#ScieloFname"/>
    </owls:EquivalentClasses>
</owls:Ontology>
```

Figura 5.7 Arquivo do modelo de integração em OWL dos modelos IBICT e SciELO

4. Após a geração dos modelos em OWL, os mesmos devem ser armazenados em alguma estrutura para futura reutilização. No protótipo foi utilizado o Microsoft SQL Server 2000 como repositório de ontologias.

5. A partir do momento em que o repositório de ontologias recebe modelos, a interface mediadora deve ter acesso ao modelo de integração para que seja executado o processo de orquestração. É esse modelo que permitirá uma visão única do ambiente.

6. Como no protótipo os modelos foram armazenados no SQL Server, só foi utilizado o filtro SQL pela interface mediador. Sendo que esse filtro terá duas solicitações, uma do modelo IBICT e outra da SciELO.

7. É importante observar que o resultado apresentado ao usuário só será estruturado quando os filtros retornarem as informações no padrão XML. Nesse caso, a interface mediadora aguarda por dois retornos para encaminhar o resultado a interface de publicação que então efetua o processamento de transformação e encaminha ao usuário. Dessa forma é finalizado o ciclo da arquitetura com relação ao protótipo.

Após a preparação da infra-estrutura, os usuários podem acessar o *site* do BASCIN para efetuar pesquisas. A seguir é apresentado o fluxo de pesquisa dos usuários.

Na Figura 5.8 é apresentado à tela inicial onde o usuário poderá efetuar a pesquisa por três campos: autor, texto e ano. No protótipo o campo texto está representando apenas o título do trabalho, mas poderia representar vários campos como resumo, palavras-chave etc. Esse processo de representação é executado no momento do mapeamento dos relacionamentos. A Figura 5.9 demonstra uma pesquisa de título que contenha a palavra “saúde”.

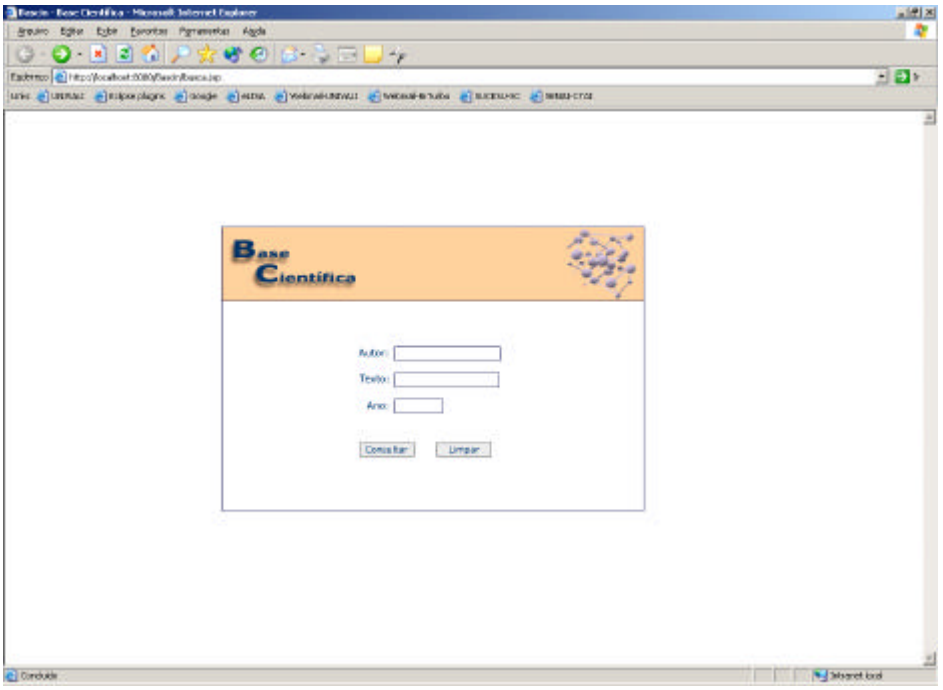

Figura 5.8 Tela inicial do BASCIN

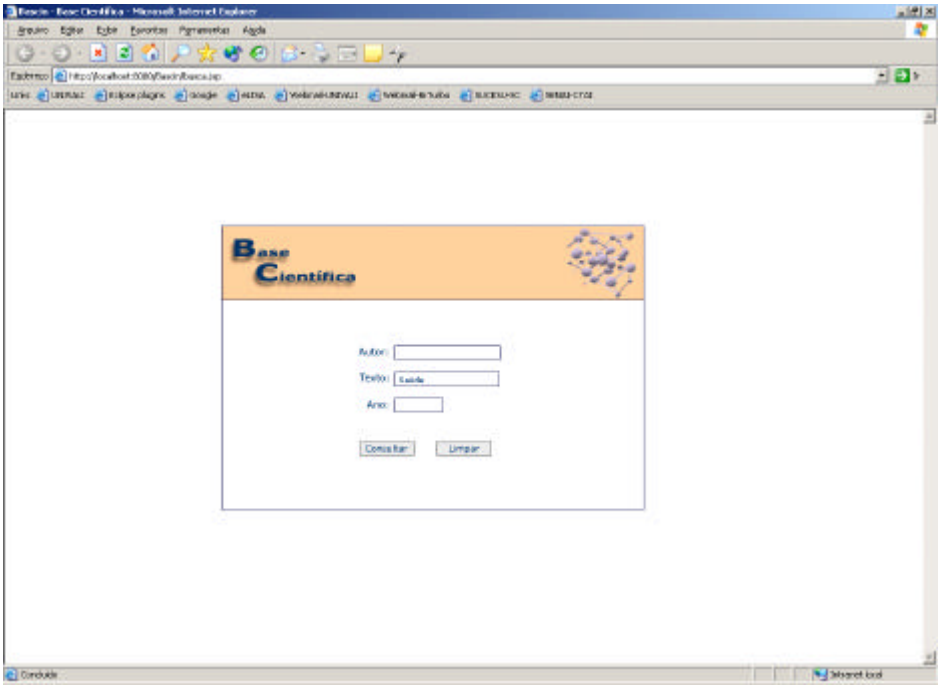

Figura 5.9 Pesquisa de títulos pela palavra “saúde”

O resultado da pesquisa encontrou produção tanto na base do IBICT quanto na base da SciELO. O resultado da pesquisa é apresentado na Figura 5.10.

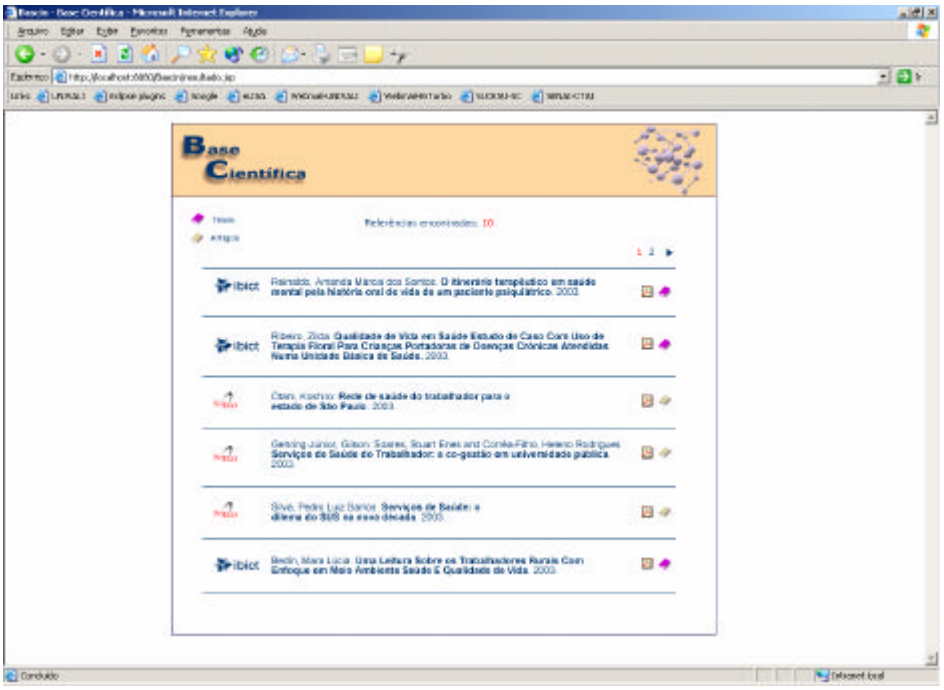

Figura 5.10 Resultado da pesquisa na BASCIN

A partir do resultado o usuário deveria acessar através dos links os trabalhos disponíveis. Como as duas bibliotecas digitais não permitem o envio de parâmetros pela URL não foi possível completar esse fluxo. De qualquer forma iremos simular essa etapa final para completar o fluxo do usuário. A Figura 5.11 apresenta informações do trabalho pesquisado, uma dissertação de mestrado. É importante lembrar que o IBICT não armazena os trabalhos na sua base, apenas os metadados dos trabalhos no padrão MTD-BR. A partir das informações da página o usuário poderá acessar diretamente a base origem do trabalho, no caso o BTD do PPGEP/UFSC como mostra a Figura 5.12.

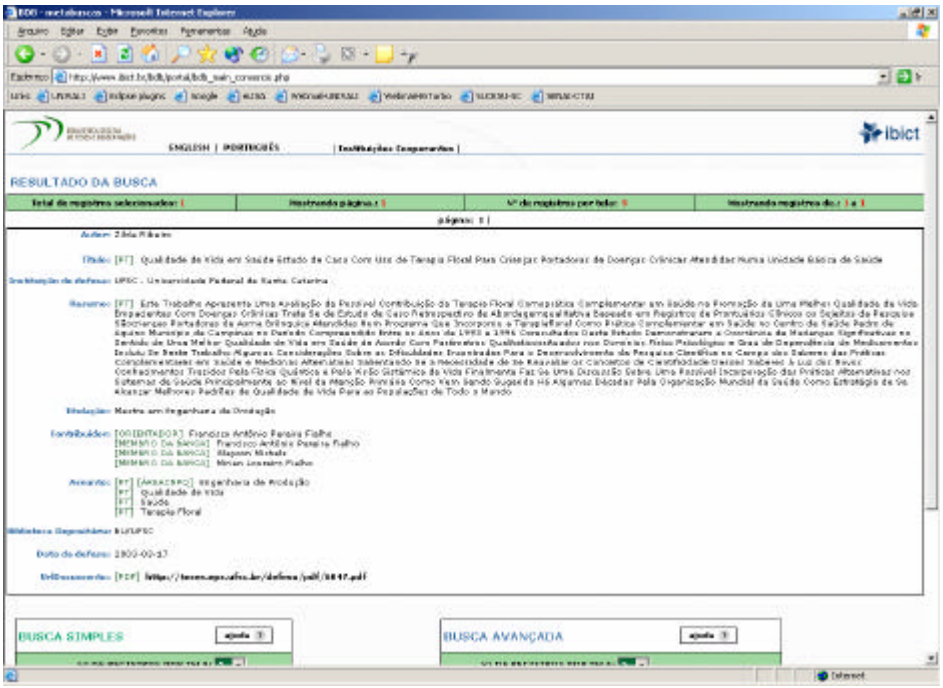

Figura 5.11 Link do texto completo no portal da BDTD do IBICT

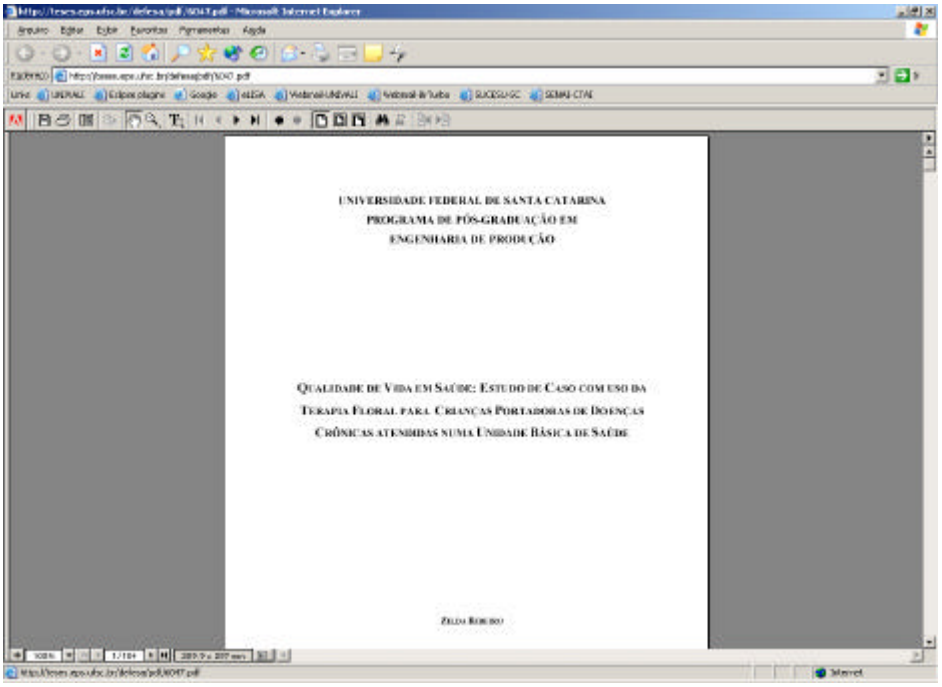

Figura 5.12 Link do texto completo no BTD do PPGEP/UFSC

Os passos seguidos no acesso aos trabalhos no portal do IBICT também deverão ser efetuados no portal da SciELO. A única diferença dos projetos é que a SciELO armazena os trabalhos e não somente os metadados na sua base conforme demonstrado nas Figuras 5.13 e 5.14.

Figura 5.13 Link do texto completo no portal da SciELO

Figura 5.14 Acesso ao texto completo da produção no portal da SciELO

5.3 Considerações finais

A abordagem de integração baseada em metadados aqui empregada reforça a importância do uso de metadados no contexto de integração de acervos científicos, conforme visto no capítulo 2, à medida que provê mecanismos que possibilitam descrever os conflitos de representação quanto à estrutura, mapeamentos, localização, e tipo de cada fonte de dados. Abordagens que fazem uso de ontologias complementam a tarefa de descrição dos recursos, uma vez que provê os mecanismos necessários para uma descrição completa da semântica acerca dos dados.

No capítulo 6 será apresentado as conclusões e principalmente os trabalhos futuros com base nesta pesquisa.

6 CONCLUSÕES E TRABALHOS FUTUROS

6.1 Conclusões

Apresentamos uma arquitetura para a integração semântica de bibliotecas digitais que procurou ser flexível o bastante para atender aos requisitos da Web Semântica. Utilizamos a abordagem de ontologias com o intuito de prover principalmente o suporte à interoperabilidade e ao consenso semântico na integração dos dados.

A XML foi adotada como a linguagem padrão na arquitetura devido suas vantagens na interoperabilidade sintática e no alto ganho de flexibilidade principalmente na apresentação dos dados em diferentes dispositivos.

As ontologias geradas na arquitetura foram representadas na linguagem OWL, que se baseia nas tecnologias RDF e XML, permitindo assim uma arquitetura extensível à implementação dos requisitos da Web Semântica.

Para validação da arquitetura foi desenvolvido um protótipo, denominado de “Base Científica”. Esse protótipo resolveu a integração semântica da Biblioteca Digital de Teses e Dissertações do IBICT e da Biblioteca Digital de Artigos da SciELO. Essa integração foi estabelecida com informações pertinentes dos autores e dos títulos dos trabalhos.

Para realizar essas relações foram geradas tabelas de correspondência entre as ontologias envolvidas, como também, regras de correspondências para a estruturação da tabela de mapeamento.

Durante o desenvolvimento da pesquisa percebeu-se que a Web Semântica representa atualmente um dos grandes desafios para a comunidade científica, pois consiste em uma integração de dados que pode ser utilizada por máquinas e humanos. Além disto, viabiliza a automação e principalmente a reutilização de dados por várias aplicações.

6.1 Trabalhos futuros

- Estudos mais aprofundados na área de gerenciamento de direitos digitais, onde essa tecnologia é mais usada para evitar que filmes, músicas e *e-books* sejam copiados. Essa tecnologia cria direitos de acesso persistentes, o que significa direitos que acompanham os objetos digitais e colocam limites sobre o encaminhamento, impressão, cópia ou alteração, e determinam quantas vezes e por quanto tempo ele pode ser visualizado.

- Estudos mais aprofundados em padrões e metodologias que garantam a persistência dos endereços eletrônicos dos recursos digitais, como:

    - URN[20] - *Uniform Resource Names*
    - PURL[21] - *Persistent URL*
    - DOI[22] - *Digital Object Identifier*

- A experiência no desenvolvimento das ontologias demonstrou a carência de ferramentas CASE no gerenciamento do repositório de

[20] http://www.w3.org/Addressing/
[21] http://purl.oclc.org/
[22] http://www.doi.org/

ontologias. Essa ferramenta a ser desenvolvida poderia auxiliar os especialistas na orquestração das ontologias.

- Implementação de *web services* que representem repositórios de ontologias e filtros de domínios distintos.

- Estudos da tecnologia de computação em grade integrada a tecnologia de *web services*.

- Desenvolvimento de agentes inteligentes para recuperação e principalmente para mapeamento de novos relacionamentos.

- Segundo (Pacheco, 2001b), a Linguagem de Marcação da Plataforma Lattes (LMPL[23]) não prevê a interoperabilidade com aplicações externas à comunidade LMPL. Sendo uma plataforma de C&T torna-se estratégico a evolução da sua forma de representação sintática para uma representação semântica, norteando uma nova realidade na interoperabilidade entre instituições de C&T.

[23] http://www.cnpq.br/lattes

7 REFERÊNCIAS BIBLIOGRÁFICAS

(BARRETO, 1999) BARRETO, C. M. **Modelo de Metadados para Descrição de Documentos Eletrônicos na Web**. Dissertação de Mestrado. Instituto Militar de Engenharia, Programa de Pós-Graduação em Engenharia de Sistemas, 1999.

(BAX, 1997) BAX, M. P. **Agentes de Interface para Bibliotecas Digitais : A Arquitetura SABiO**. In: Seminário sobre automação em bibliotecas e centros de documentação, 6., 1997, Águas de Lindóia. Anais... Águas de Lindóia : UNIVAP, 1997.

(BECHHOFER et al., 2000) BECHHOFER, S. et al. ***An Informal Description of Standard OIL and Instance OIL***, 2000. Disponível em <http://www.ontoknowledge.org/oil/>. Acessado em 27 ago. 2003.

(BERGAMASCHI et al., 1999) BERGAMASCHI, S. et al. ***Semantic Integration of Semistructured and Structured Data Sources***. SIGMOD Record, 1999.

(BERNERS – LEE et al., 2001a) BERNERS – LEE, T. et al. ***Semantic Web Development Proposal*****,** 2001. Disponível em: <http://www.w3c.org/2001/sw/>. Acesso em 10 ago. 2003.

(BERNERS – LEE et al., 2001b) BERNERS – LEE, T. et al. ***The Semantic Web***. Scientifc American, 2001.

(BREUKER, 1994) BREUKER, J., VAN DE VELDE, W. ***CommonKADS Library for Expertise Modelling***, IOS Press, 1994.

(CALVANESE et al., 1998) CALVANESE, D. *et a*l. ***Information Integration: Conceptual Modeling and Reasoning Support****.* Conference on Cooperative Information Systems, 1998.

(CASTOLDI, 2003) CASTOLDI, A. V. **Uma Ontologia para Enlaces de Unidades de Informação em Plataformas de Governo Eletrônico.** Dissertação de Mestrado. Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção. Florianópolis, 2003.

(CLANCEY, 1993) CLANCEY, W. J. ***The Knowledge Level Reinterpreted: Modelling Socio-Technical Systems****.* International Journal of Intelligent Systems, 1993.

(CHANDRASEKARAN, 1997) CHANDRASEKARAN, B., JOSEPHSON, J. R. ***The Ontology of Tasks and Methods*.** Stanford University, California, 1997.

(CWM, 2003) ***OMG Common Warehouse Metamodel (CWM) Specification***. Disponível em: <http://www.omg.org/cwm/>, 2003. Acessado em 28 ago. 2003.

(DACONTA, 2003) DACONTA, M.C., OBRST, L.J., SMITH, K.T. ***The Semantic Web***. Wiley Publishing. Indianapolis, 2003.

(DAML+OIL, 2000) <http://www.daml.org/2000/12/daml+oil-index>.

(FALBO, 1998) FALBO R. A. **Integração de Conhecimento em um Ambiente de Engenharia de Software.** Tese de Doutorado. Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Engenharia de Sistemas e Computação. Rio de Janeiro, 1998.

(FENSEL, 2000) FENSEL D**.** et al**. *OIL in a Nutshell*** In: Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference, 2000.

(FÉRNANDEZ, 1997) FÉRNANDEZ, M., GÓMEZ-PÉREZ, A., JURISTO, N. ***METHONTOLOGY: From Ontological Art Towards Ontological Engineering***. Ontological Engineering - Working Notes, Stanford University, California, 1997.

(GRUBER, 1992) GRUBER, T. R. ***Ontolingua: A Mechanism to Support Portable Ontologies, version 3.0***. Relatório Técnico. Knowledge Systems Laboratory, Stanford University, California, 1992.

(GRUBER, 1995) GRUBER, T. R. ***Toward Principles for the Design of Ontologies used for Knowledge Sharing*.** Int. J. Human-Computer Studies, v. 43, n. 5/6, 1995.

(GUARINO, 1997)GUARINO, N. ***Understanding, Building and Using Ontologies***. Int. Journal Human-Computer Studies, v. 45, n. 2/3, 1997.

(GUARINO, 1998a) GUARINO, N., WELTY, C. ***Conceptual Modeling and Ontological Analysis***. LADSEB-CNR, Padova, 1998.

(GUARINO, 1998b) GUARINO, N. ***Formal Ontologies and Information Systems***. In: First International Conference, 1., 1998, Trento, Itália. Anais... Trento: IOS Press, 1998.

(GUHA, 1997) GUHA, R., BRAY, T. ***Meta Content Framework Using XML***, 1997. Disponível em: <http://www.w3.org/TR/NOTE-MCF-XML-970606>. Acessado em 05 set. 2003.

(GUNTHER, 1997) GUNTHER, O., VOISARD, A. ***Metadata in Geographic and Environmental Data Management***. In: W. Klas, e A. Sheth, Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data. McGraw Hill, 1997.

(GUNTHER, 1998) GUNTHER, O. ***Environmental Information Systems***. Springer-Verlag, 1998.

(GRUTTER, 2001) GRÜTTER, R., EIKEMEIER, C. ***Development of a Simple Ontology Definition Language (SontoDL) and its Application to a Medical Information Service on the World Wide Web***. The First Semantic Web Working Symposium, California, 2001.

(HASSELBRING, 2000) HASSELBRING, W. ***Information System Integration***. Communications of the ACM, vol.43, n.6, June 2000.

(HEFLIN, 2000) HEFLIN, J., HENDLER, J. ***Semantic Interoperability on the Web***. Extreme Markup Languages, 2000.

(HORROCKS, 2000) HORROCKS, I., FENSEL D., BROEKSTRA J., DECKER S., et al. ***The Ontology Layer OIL***, 2000. Disponível em <http://www.ontoknowledge.org/oil//TR/oil.long.html>. Acessado em 11 ago. 2003.

(HULL, 1997) HULL, R. ***Managing Semantic Heterogeneity in Databases: A Theoretical Perspective***. ACM PODS, 1997. Disponível em <http://www-db.research.bell-labs.com/user/hull/pods97-tutorial.html>. Acessado em 20 ago.2003.

(KARP, 1999) KARP P.D., CHAUDHRI V.K., THOMER J. X.: ***An XML-Based Ontology Exchange Language***, Tech. Rep., Versions 0.3, 1999.

(KENT, 1989) KENT, W. ***The Many Forms of a Single Fact***. San Francisco: Proc. IEEE Compcon, 1989. Disponível em: <http://www.bkent.net/Doc/manyform.htm>. Arquivo consultado em 28 ago. 2003.

(KERHERVÉ, 1997) KERHERVÉ, B. ***Models for Metadata or Metamodels for Data?***. Second IEEE Metadata Conference, Silver Spring, Maryland, Disponível em: <http://citeseer.nj.nec.com/kerherve97model.html>. Acessado em 22 ago. 2003.

(KRAUSKOPF, 1996) KRAUSKOPF, T. et al. ***PICS Label Distribution Label Syntax and Communication Protocols Version 1.1***. W3C (World-Wide Web Consortium) Disponível em: <http://www.w3.org/TR/REC-PICS-labels>. Acessado em 02 set. 2003.

(KRISHNAMURTHY, 1991) KRISHNAMURTHY, R., LITWIN, W., KENT, W. ***Language Features for Interoperability of Data Base with Schematic Discrepancies***. ACM, 1991.

(LAGOZE, 1996) LAGOZE, C. et al. ***The Warwick Framework - A Container Architecture for Aggregating Sets of Metadata.*** Disponível em: <http://www.dlib.org/dlib/july96/lagoze/07lagoze.html>. Acessado em 19 ago. 2003.

(LASSILA, 1999) LASSILA, Ora, SWICK, Ralph R. ***Resource Description Framework (RDF) Model and Syntax Specification***, Disponível em: <http://www.w3c.org/TR/1999/REC-rdf-syntax-19990222>. Acessado em 26 ago. 2003.

(LEVACOV, 1997) LEVACOV M. **Bibliotecas Virtuais : (R)evolução?..** Ci. Inf, Brasília, DF, v. 26, n.2, p.125-135, maio/ago. 1997.

(LEVY, 1995) LEVY, D. M., MARSHAL, C.C.. ***Going Digital: A Look at Assumptions Underlying Digital Libraries*.** Communications of the ACM v38, n4, 1995.

(LIMA, 2001) Lima , T. S. **Entrelaçamentos Semânticos Modelando a Integração da Informação na Web Semântica**. Tese de Doutorado, Universidade do Estado de São Paulo, Programa de Pós-Graduação em Ciência da Computação, São Carlos, 2001.

(LUKE, 2000) LUKE, S., HEFLIN, J. ***SHOE 1.01. Proposed Specification***, Disponível em: <http://www.cs.umd.edu/projects/plus/SHOE/spec.html>. Acessado em 14 ago. 2003.

(MARTIN, 2001) MARTIN, D.; BIRBECK, M.; KAY, M.; LOESGEN, B.; PINNOCK, J.; LIVINGSTONE, S.; WILLIAMS, K.; ANDERSON, R.; MOHR, S.; BALILES, D.; PEAT, B.; OZU, N. ***Professional XML***. Ciência Moderna Ltda, 2001.

(MOURA, 1998) MOURA, A.,CAMPOS, M. L. M., BARRETO, C. M. ***A Survey on Metadata for Describing and Retrieving Internet Resources***. World Wide Web Journal, Baltzer Science Publishers BV, 1998.

(NOY, 1997) NOY, N. F., HAFNER C.D. ***The State of Art in Ontology Design: A Survey and Comparative Review***. AI Magazine, 1997.

(O'LEARY, 1997) O'LEARY, D. E. ***Impediments in the Use of Explicit Ontologies for KBS Development***. Int. J. Human-Computer Studies, v. 46, n. 2/3, 1997.

(OIM, 1999) ***Meta Data Coalition. Open Information Model (OIM) version 1.1 Proposal***, Disponível em: http://www.MDCinfo.com. Acessado em 02 ago. 2003).

(OLIVEIRA, 2002) OLIVEIRA, Douglas H. **Introdução a XML e suas Aplicações**, 2002.

(PACHECO, 2001a) PACHECO, R. C. S; KERN, V. M. **Transparência e Gestão do Conhecimento por Meio de um Banco de Teses e Dissertações: A Experiência do PPGEP/UFSC.** Ci. Inf., Brasília, v. 30, n. 3, p. 64-72, set./dez. 2001

(PACHECO, 2001b) PACHECO, R. C. S; KERN, V. M. **Uma Ontologia Comum para a Integração de Bases de Informações e Conhecimento sobre Ciência e Tecnologia**. Ci. Inf., Brasília, v. 30, n. 3, p. 56-63, set./dez. 2001

(PADILHA, 2003) PADILHA, P. **Extrator de Metadados**. Trabalho de Conclusão de Curso, Faculdades Integrada Facvest, Curso de Ciências da Computação, Lages, 2003.

(PICS, 1996) ***PICS: Internet Access Controls Without Censorship***, Communications of the ACM, Vol.39, N.10, October 1996.

(RDFS, 2000) ***Resource Description Framework (RDF) Schemas***. W3C Candidate Recommendation, Disponível em: <http://www.w3.org/TR/rdf-schema/>. Acessado em 10 ago. 2003).

(RUSSEL, 1995) RUSSEL, S., NORVIG, P. ***Artificial Intelligence: A Modern Approach***. Prentice-Hall, 1995.

(SAUNDERS, 1995) Saunders, L. M. ***Transforming Acquitions to Support Virtual Libraries. Information Technology and Libraries***, Vol. 14, N. 1, March 1995.

(SIMON, 1998) SIMON, E., TOMASIC, A., GALHARDAS, H. ***A Framework for Classifying Scientific Metadata***. American Association for Artificial Intelligence, 1998.

(SMITH, 2000) SMITH, B. ***Ontology: Philosophical and Computational***. Disponível em: <http://wings.buffalo.edu/philosophy/faculty/smith/articles/ontologies.htm>, Acessado em: 03 set. 2003.

(SMITH, 2003) SMITH, Michael K; VOLZ, Raphael; McGUINESS, Debora. ***Web Ontology Language (OWL) Guide Version 1.0***, W3C Working Draft, Disponível em: <http://www.w3.org/TR/2002/WD-owl-guide-20021104>. Acesso em 22 de ago. 2003.

(SOAP, 2003) <http://www.w3c.org/TR/soap12-part0/>, Acessado em 15 set. 2003

(SOWA, 2000) SOWA, J. F. ***Ontology, Metadata, and Semiotics***. ICCS'2000 in Darmstadt, Germany, August, 2000. Disponível em: <http://users.bestweb.net/~sowa/peirce/ontometa.htm>. Acessado em 17 ago. 2003.

(SWOBODA, 1999) SWOBODA, W. et al. ***The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany***. IEEE, 1999. Disponível em <http://www.computer.org/proceedings/meta/1999/papers/45/wswoboda.html>. Acessado em 25 ago. 2003.

(UDDI, 2002) <http://uddi.org/pubs/uddi-v3.00-published-20020719.htm>, Acessado em 20 set. 2003

(USCHOLD, 1995) USCHOLD, M., KING, M. ***Towards a Methodology for Building Ontologies***. In: Workshop on Basic Ontological Issues in Knowledge Sharing,1995.

(USCHOLD, 1996) USCHOLD, M., GRUNINGER M. ***Ontologies: Principles, Methods and Applications***. The Knowledge Engineering Review, v. 11, n. 2, p. 93-136, 1996.

(XMI, 2003) ***XML Metadata Interchange (XMI) Specification*****, Verson 2.0**, 2003. Disponível em: <http://www.omg.org/cgi-bin/doc?formal/2003-05-02>. Acessado em 24 ago. 2003.

(Web Services, 2002) <http://www.w3c.org/2002/ws/>, Acessado em 14 ago. 2003.

(WEIBEL, 1997) Weibel, S., Miller,r. ***Dublin Core Metadata Element Set Reference Description Office of Research***. OCLC Online Computer Library Center, Inc. Web Site: http://www.purl.org/ . 1997.

(WEIBEL, 1999) WEIBEL, S. ***The State of the Dublin Core Metadata Initiative***. D-Lib Magazine, April 1999.

(WOODRIDGE, 1999) WOODRIDGE, M. ***Intelligent Agents***. In G. Weiss, Multiagent Systems, The MIT Press, April 1999.

(WOOLF, 1981) WOOLF H. B. ***Webster's New Collegiate Dictionary***. Springfield. Mass: G&C, Merriam, 1981.