UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

INTEGRAÇÃO DE DADOS BASEADA EM ONTOLOGIA

Gustavo Ferreira Afonso

SÃO CARLOS
2008

Verso da folha de rosto (Ficha Catalográfica)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

INTEGRAÇÃO DE DADOS BASEADA EM ONTOLOGIA

Gustavo Ferreira Afonso

Dissertação apresentada ao programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para obtenção do Título de Mestre em Ciência da Computação.
Orientador: Prof. Dr. Mauro Biajiz
Co-Orientadora: Profa. Dra. Marilde Terezinha Prado Santos

SÃO CARLOS
2008

Folha de aprovação

Aos meus pais.

AGRADECIMENTOS

Ao que criou, o Infinito e Perfeito, permitindo a existência.

Aos meus pais, a quem devo minha criação e inspiração.

Ao Professor Mauro e Professora Marilde, pela orientação e dedicação empregados, permitindo a concretização do trabalho.

Aos Professores Francisco Dupas, José Tundisi e Takako Tundisi, pelo ensino e auxílio, encaminhadores nesta jornada acadêmica.

À Izabel Barbelli (Bel), pela motivação e assessoria filosófica.

À Cristiane Yaguinuma, pela assessoria técnica, responsável pelo módulo de expansão da consulta.

Aos colegas e amigos do IIE, que sempre colaboraram com contentamento.

Aos colegas e amigos do DC, mestres honorários, que desde o início estiveram presentes nas superações dos desafios.

Aos Professores e Funcionários do DC, que participaram do desenvolvimento desta caminhada.

À CAPES, pelo apoio financeiro e ao IIE, pelo apoio à pesquisa e por disponibilizar dados e conhecimentos.

Muito obrigado a todos que fazem o universo conspirar a nosso favor.

RESUMO

O estudo sobre integração de dados não é recente, mas tampouco está esgotado. Sua necessidade destaca-se cada vez mais devido à busca por informações em um ambiente geral com crescente disponibilidade de dados, provenientes de fontes distribuídas e heterogêneas. Uma abordagem empregada com êxito, principalmente para solucionar os problemas de heterogeneidades das fontes, é o emprego de ontologia, como base para os sistemas de integração de dados. Com objetivo de empregar esse conceito, este trabalho descreve o desenvolvimento de um sistema de integração de dados baseado em ontologia, denominado DISFOQuE, que tem como motivação seu uso no domínio de análises de bacias hidrográficas.

Palavras-chave: integração de dados, ontologia, sistema de integração de dados, heterogeneidade, bacias hidrográficas.

ABSTRACT

The study on data integration is not new but neither is exhausted. Its need is increasingly clear due to the search for information in a general environment with increasing availability of data proceeding from distributed and heterogeneous sources. An approach used successfully, especially to solve the problems of heterogeneity of sources, is the use of ontology, as the basis for the data integration systems. In order to employ this concept, this work describes the development of a data integration system based on ontology, DISFOQuE, which has its use as a motivation in the field of watersheds analysis.

Keywords: data integration, ontology, data integration system, heterogeneity, watersheds.

LISTA DE FIGURAS

LISTA DE TABELAS E QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BD	Banco de Dados
DAML	*DARPA Agent Markup Language*
DISFOQuE	*Data Integration System using Fuzzy Ontology-based Query Expansion*
DW	*Data Warehouse*
FOQuE	*Fuzzy Ontology-based Query Expansion*
IIE	Instituto Internacional de Ecologia
IIEGA	Instituto Internacional de Ecologia e Gerenciamento Ambiental
KRAFT	*Knowledge Reuse And Rusion/Transformation*
OIL	*Ontology Inference Layer*
OKBC	*Open Knowledge Base Connectivity*
OLAP	*On-Line Analytical Processing*
OQL	*Object Query Language*
OWL	*Web Ontology Language*
RDF	*Resource Description Framework*
RUP	*Rational Unified Process*
SGBD	Sistema Gerenciador de Banco de Dados
SID	Sistema de Integração de Dados
SIG	Sistema de Informação Geográfica
SIMS	*Services and Information Management for decisions Systens*
SQL	*Structured Query Language*
TAMBIS	*Transparent Access to Multiple Bioinformatics Information Sources*
W3C	*World Wide Web Consortium*
WWW	*World Wide Web*
XML	*EXtensible Markup Language*

SUMÁRIO

1. Introdução

Integração de dados é um campo de pesquisa da comunidade de banco de dados que cresce rapidamente há três décadas, desde a criação e inclusão dos sistemas de banco de dados nas empresas [54]. Cada vez mais se destaca em meio à necessidade de buscas de informações em um ambiente de crescente disponibilidade de dados em fontes distribuídas e heterogêneas. Nesse desenvolvimento, surgem várias propostas para solucionar os problemas nessa área. Dentre essas propostas, há a abordagem baseada no uso da ontologia com o intuito de prover semântica aos dados.

Motivados a aplicar esse conceito de integração de dados baseado em ontologia em um contexto do mundo real, utiliza-se nesta pesquisa o domínio encontrado dentro do IIE (Instituto Internacional de Ecologia) e IIEGA (Instituto Internacional de Ecologia e Gerenciamento Ambiental) de São Carlos – SP. Nesses institutos, projetos de pesquisas e consultorias com escopo na análise de bacias hidrográficas são referências dentro e fora do Brasil. Dessa forma, com objetivo de desenvolver um SID (Sistema de Integração de Dados) com motivação no domínio de análises de bacias hidrográficas provido pelos institutos IIE/IIEGA, gera-se o trabalho de mestrado descrito nesta dissertação.

1.1. Motivação

O domínio de análises de bacias hidrográficas é bastante oportuno à pesquisa de integração de dados baseada em ontologia, uma vez que apresenta várias fontes de dados isoladas e heterogêneas possíveis de serem integradas. A integração dessas fontes de dados, provenientes de diferentes trabalhos, apresenta a possibilidade de respostas a consultas mais elaboradas e/ou análises a quantidades de dados maiores e mais expressivos. Por exemplo, uma consulta para obter os dados de fósforo total durante os anos de 1990 a 2007 no rio Tietê. Possivelmente esses dados então distribuídos entre várias fontes heterogêneas, e mesmo que seja possível responder a essa consulta pesquisando apenas uma única fonte, a resposta se tornaria mais significativa se uma quantidade maior de dados fosse enviada. Em um outro exemplo, a consulta para obter os dados de nitrogênio total em coletas onde a temperatura do ar estivesse abaixo dos 15 °C é formulada. Suponha que existam várias fontes, mas nenhuma delas com ambos os dados, de nitrogênio total e de temperatura do ar, inseridos; e considere também algumas fontes contendo dados de nitrogênio total e outras fontes contendo dados da

temperatura do ar em determinadas regiões. Então, essa consulta só poderá ser respondida se houver a integração dos dados dessas diferentes fontes.

A integração de dados traz desafios, sendo a heterogeneidade das fontes um dos principais. Para tratar essa questão, o uso de ontologia vem sendo empregado, constituindo uma visão semântica dos termos do domínio a ser integrado. Como vantagem dessa visão semântica à integração de dados, a ontologia provê o entendimento comum e não ambíguo dos termos e conceitos, servindo como visão única e homogênea do domínio.

1.2. Objetivo

O objetivo primário do presente trabalho é voltado ao desenvolvimento de um Sistema de Integração de Dados (SID) baseado em ontologia, motivado ao uso no domínio de bacias hidrográficas. Dessa forma, funções especiais são criadas para solucionar problemas característicos desse domínio. Como exemplo, um Banco de Dados Geográficos que identifica a localização espacial de coordenadas geográficas de pontos quaisquer, entre outras funções relacionadas a informações geográficas. Existem outras utilidades desempenhadas pelo sistema motivadas por esse domínio específico, como o tratamento das unidades de medidas necessárias às consultas realizadas. Além dessas funções voltadas ao domínio de bacias hidrográficas, a utilização do módulo de expansão da consulta, criado no trabalho de Yaguinuma [11], concede às consultas do sistema a possibilidade de uma expansão semântica, utilizando-se da lógica difusa (*fuzzy*) presentes na ontologia. O uso desse módulo dá origem ao nome do SID desenvolvido, denominando-o com a sigla DISFOQuE (*Data Integration System using Fuzzy Ontology-based Query Expansion*). No trabalho de Yaguinuma, o desenvolvimento de uma meta-ontologia possibilita a utilização de regras difusas de similaridade, proximidade todo-parte e transitividade, empregadas na construção de ontologias para o SID apresentado aqui.

Entretanto, o uso do Sistema de Integração de Dados desenvolvido é possível e viável em outros domínios de aplicação. Apesar das funções especiais, motivadas pelo domínio de bacias hidrográficas, possivelmente não serem empregadas em outros domínios, a utilização do SID nestes não é prejudicada. O mesmo SID é flexível para desempenhar suas funções básicas em um domínio diferente, necessitando de uma ontologia que represente este novo domínio e de mapeamentos entre as fontes deste domínio e a nova ontologia.

1.3. Organização

A estrutura desta dissertação encontra-se organizada da seguinte forma:

- O capítulo 2 descreve o domínio do caso de estudo, análises de bacias hidrográficas, de forma sucinta e multidisciplinar. Inicia-se descrevendo o domínio de forma geral, seguido da descrição dos seus principais dados e as dificuldades em se trabalhar com estes dados;
- O capítulo 3 aborda o tema ontologia, explicando o conceito, a constituição e os tipos de ontologia. Prossegue com uma discussão mais técnica, sobre metodologias de construção e linguagens de implementação de ontologia;
- O capítulo 4 trata de forma geral o tema integração de dados. Emite a motivação e definição do assunto e prossegue comentando os desafios e abordagens da integração de dados. Finaliza com um estudo do uso de ontologia nos Sistemas de Integração de Dados, tema de interesse mais específico neste trabalho;
- O capítulo 5 apresenta o Sistema de Integração de Dados desenvolvido, o DISFOQuE. Expõe as abordagens utilizadas na construção, as partes que o constituem e seu fluxo de funcionamento. Detalha o mapeamento realizado entre ontologia e fontes, as tecnologias empregadas, e finaliza apontando as soluções realizadas aos desafios da integração de dados;
- O capítulo 6 detalha como foi realizado o caso de estudo do domínio de bacias hidrográficas, motivador ao desenvolvimento deste trabalho. Apresenta a construção da ontologia, detalha, em um exemplo, o funcionamento do sistema, mostra o emprego do DISFOQuE em outros domínios e comenta o desempenho deste sistema em relação ao tempo de resposta;
- O capítulo 7 finaliza este documento, manifestando as conclusões gerais, discutindo os trabalhos relacionados e resultados atingidos. Também são listadas as propostas para trabalhos futuros.

2. O domínio do caso de estudo

O SID desenvolvido, denominado DISFOQuE, teve especial atenção para integrar dados de bacias hidrográficas. Tal domínio foi escolhido não somente para uma descrição mais detalhada do DISFOQuE, mas também como motivação para construção deste sistema e assim, adequação do mesmo às características específicas do domínio de bacias hidrográficas.

Os conhecimentos e fontes de dados necessários para esse caso de estudo envolvendo bacias hidrográficas são, em grande parte, provenientes do IIE/IIEGA de São Carlos-SP. Muitos projetos relacionados às análises de bacias hidrográficas, com coletas de uma infinidade de dados de vários tipos, são constantemente realizados nesse instituto. Os dados desses projetos são armazenados na maioria dos casos em documentos textos ou planilhas eletrônicas e uma pequena parte é armazenada em bancos de dados estruturados. Pretende-se, com o trabalho finalizado, realizar aplicações mais intensas do sistema desenvolvido em fontes de dados desse instituto.

Este capítulo descreve o domínio de bacias hidrográficas de maneira sucinta e multidisciplinar. Primeiro a seção 2.1 o descreve de forma geral, sem nenhum relacionamento com a ciência da computação e áreas afins, o domínio estudado, com seus conceitos básicos e necessários para o entendimento de partes deste trabalho. Posteriormente na seção 2.2, examinam-se os dados importantes do domínio, envolvendo-os a conceitos computacionais de banco de dados. Então na seção 2.3, em uma discussão multidisciplinar, são apresentadas as dificuldades em se trabalhar com os dados dentro desse domínio estudado. Finalizando, na seção 2.4 são dadas as conclusões do capítulo.

2.1. Descrição geral do domínio estudado - bacias hidrográficas

Pela definição do dicionário Aurélio [1]: “bacia hidrográfica, bacia de drenagem ou bacia fluvial é o conjunto das terras drenadas por um rio e por seus afluentes”, ou seja, região limitada pelos divisores de água, formada por um rio principal e seus afluentes, como é mostrado na Figura 2.1. Na legislação brasileira [2]: “a bacia hidrográfica é a unidade territorial para implementação da Política Nacional de Recursos Hídricos e atuação do Sistema Nacional de Gerenciamento de Recursos Hídricos”.

Na primeira definição, bacias hidrográficas são vistas como unidades naturais da paisagem e não relacionadas com limites políticos. Mesmo que ao se fazer uma delimitação legal, muitas vezes feita por comitês de bacias hidrográficas, se leve em conta o aspecto do relevo geográfico, isso não está explícito na definição da legislação brasileira. E é o que ocorre em muitos casos de delimitação para estudos de bacias hidrográficas, no qual os aspectos políticos, como a divisão municipal, são o fator determinante na delimitação das bacias. Mas em qualquer caso a água que corre nos rios, ribeirões, lagos e represas, é o elemento formador da bacia hidrográfica.

A Figura 2.1 apresenta uma ilustração de uma bacia hidrográfica, delimitada pela curva de nível mais alta do terreno (traço delimitando limite maior e externo), formada pelo rio principal e seus afluentes. Em um desses afluentes é mostrado o limite de sua área de drenagem (traço delimitando limite menor e interno), formando-se uma sub-bacia da bacia principal. São destacadas algumas nascentes dos rios, ou seja, o lugar onde brota essas correntes de água. Os pontos *A*, *B* e *C* foram colocados para exemplificar o significado das palavras montante e jusante. O ponto *A* está à montante em relação ao ponto *B*, correndo a água no sentido de *A* para *B*. Já o ponto *C* está à jusante em relação ao mesmo ponto *B*, correndo a água no sentido de B para C.

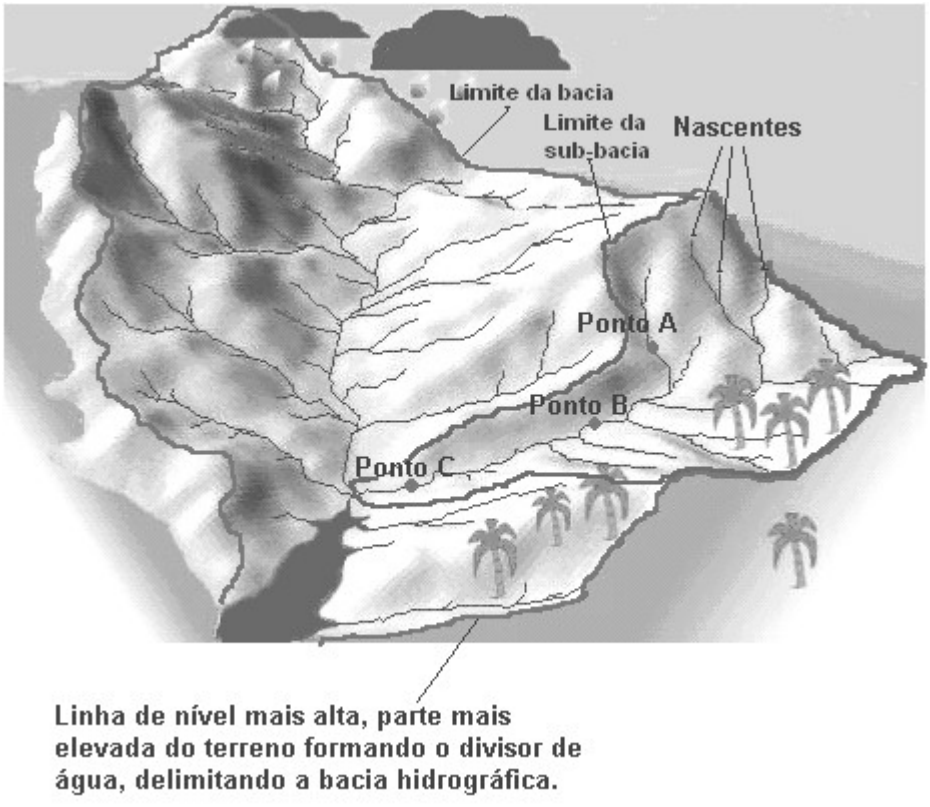


Figura 2.1 – A bacia hidrográfica. (adaptado de [3])

A Figura 2.1 destaca conceitos importantes no estudo de bacias hidrográficas, mostrando suas relações e dependências, o que introduz a idéia de que a bacia hidrográfica não é algo isolado e tampouco relacionado apenas aos rios formadores da bacia.

Como apresentado nessa figura, uma bacia maior pode ser dividida em regiões menores, sendo assim chamadas de sub-bacias desta bacia maior. Com isso é possível definir melhor as áreas de estudos, com regiões menores para um melhor detalhamento, uma vez que é difícil gerenciar as grandes áreas das bacias. Já essas sub-bacias incluem microbacias, que devem ser consideradas como unidades básicas de manejo, visto que alterações feitas em qualquer ponto da mesma podem acarretar melhorias ou comprometimento da qualidade e quantidade da água disponível. Tanto a sub-bacia quanto as microbacias são delimitadas pelos divisores de água, definidos pelas curvas de níveis do relevo do terreno. No entanto, assim como no caso das bacias, em alguns estudos outras questões podem ser consideradas, como por exemplo, os limites municipais.

Mesmo sendo a água o elemento principal de uma bacia hidrográfica, um estudo sobre este domínio não se restringe apenas em avaliar os componentes naturais da água formadora desta bacia. As características biogeofísicas são influentes na formação e afetam a bacia hidrográfica, alterando a composição e qualidade de seu elemento principal, a água. Assim, um estudo englobando vários fatores, como o uso e ocupação da área, o levantamento dos atributos físicos, composição do solo, relevo, biodiversidade aquática, atributos químicos e físicos da própria água etc, se torna mais eficiente; formando sistemas hidrológicos e ecológicos mais coerentes com a dinâmica da bacia hidrográfica. Portanto, a seção seguinte trata desses fatores relevantes ao estudo de bacias hidrográficas.

2.2. Dados pertinentes ao domínio estudado

Uma variedade grande de parâmetros compõe os fatores influentes e avaliadores das bacias hidrográficas. Na maioria das vezes a qualidade da análise das bacias está relacionada à quantidade de registros captados para tal estudo. Com isso, se torna desejável o maior número de dados concernentes ao estudo em questão. Porém, devido à grande variedade de parâmetros relacionados ao estudo, se torna quase impossível em uma análise o uso de todos os dados que são considerados pertinentes. Assim, este documento abstrai o domínio, comentando os parâmetros geralmente avaliados nesses estudos, de forma breve e organizados nos seguintes grupos:

- Parâmetros de uso e ocupação do solo: referentes ao que está sobre o solo. Qual o tipo de vegetação (incluindo pastagem, desmatamento e plantações) ou edificação e qual a população da área. Tipicamente esses parâmetros são trabalhados em um SIG (Sistema de Informação Geográfica). Fotos aéreas ou de satélites são utilizadas na composição de imagens utilizadas nas classificações do uso do solo, juntamente às coletas em campo que dão as amostras necessárias para estas classificações. Os dados de ocupação são obtidos pelos censos, algumas vezes mais detalhados por áreas dos municípios pertinentes;
- Parâmetros físicos, químicos e limnológicos da água: caracterizam a água conforme sua composição e estado. Estão relacionados entre si e com o sistema, como a profundidade em que são coletados. A obtenção desses parâmetros é feita por coletas de água em locais pontuais, com anotação das coordenadas geográficas, assim como da profundidade e/ou altitude em que foram coletados. Os locais dessas coletas pontuais devem ser escolhidos baseados em uma lógica para representar de forma mais ampla e precisa a bacia hidrográfica analisada. Posterior às coletas, estas amostras são analisadas geralmente em laboratório onde os dados dos parâmetros são obtidos;
- Parâmetros das espécies biológicas na água: organismos, incluindo os microorganismos, que estão presentes na água. A obtenção desses parâmetros também é relacionada à sua localização geográfica. Em um estudo no qual são feitas tantas as coletas de parâmetros físicos, químicos ou limnológicos quanto as coletas de espécies biológicas, é sensato fazê-las juntas, ou seja, no mesmo local e horário se faz a coleta de todos os parâmetros, podendo assim relacioná-las de forma mais coerente;
- Parâmetros que caracterizam as condições do trecho hidrográfico: medidas das dimensões e das características hidrológicas de uma área ou trecho hidrográfico. Em muitos casos os parâmetros hidrológicos são obtidos por estações fixas de coletas. Existe uma dificuldade maior em se obter alguns desses dados no local exato em que se obteve os dados dos dois tipos de parâmetros anteriores e muitas vezes não são coletados em um mesmo local e tempo. Nesse caso, para relacionar todos esses parâmetros é necessário uma expansão da área de coleta dos dados hidrológicos.

Há duas características comuns entre todos esses diversos parâmetros do domínio, sendo elas a localização espacial e a localização temporal. No primeiro caso, a localização espacial dos parâmetros permite analisar uma dependência existente entre os dados de regiões distintas, além de comparações entre diferentes áreas e construções de

mapas, que permitem uma visualização mais clara e facilita ações de gerenciamento e preservação das bacias hidrográficas. O caso da localização temporal é fundamental, uma vez que o domínio como um todo está intimamente ligado a fatores climáticos, muitas vezes característicos de determinados períodos anuais, além de possibilitar as análises históricas.

A Figura 2.2 é um exemplo de modelo de dados para o domínio de bacias hidrográficas, apresentando os parâmetros das análises e seus relacionamentos. Nessa figura observa-se que a entidade *coleta*, que é uma entidade central para o modelo de dados, possui os atributos *latitude* e *longitude* para sua localização geográfica e o atributo *data_coleta* para identificar a época da coleta. Esse modelo também estrutura os dados de uso e ocupação do solo, o que facilita um estudo mais completo das bacias hidrográficas segundo o conceito de que estas não são dependentes apenas dos parâmetros coletados nas águas de seus rios.

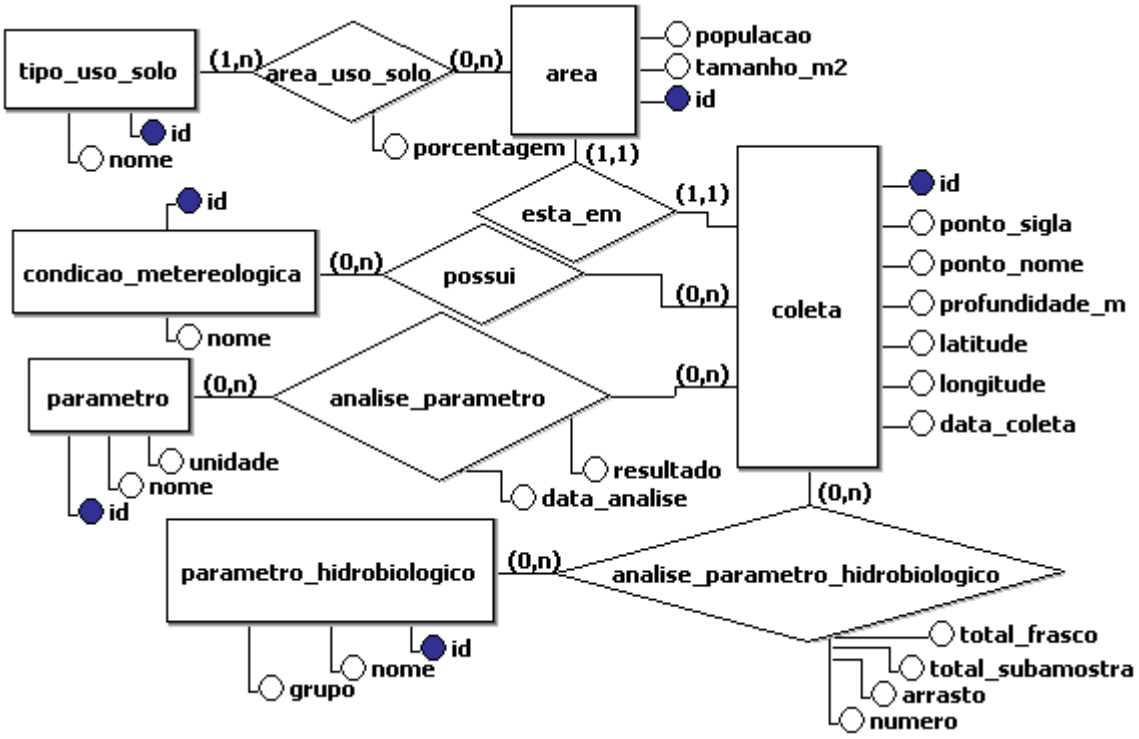


Figura 2.2 – Um modelo de dados pertinentes ao domínio estudado.

Uma vez visto os dados pertinentes ao domínio de bacias hidrográficas, seus relacionamentos e características de obtenção, a seção seguinte discute as dificuldades em se manipular tais dados.

2.3. Dificuldades em se trabalhar com os dados do domínio estudado

Para iniciar uma discussão sobre as dificuldades em se trabalhar com os dados do domínio, a Figura 2.3 apresenta um exemplo típico de como os dados do domínio são trabalhados, desde sua obtenção, tratamento, armazenamento e utilização.

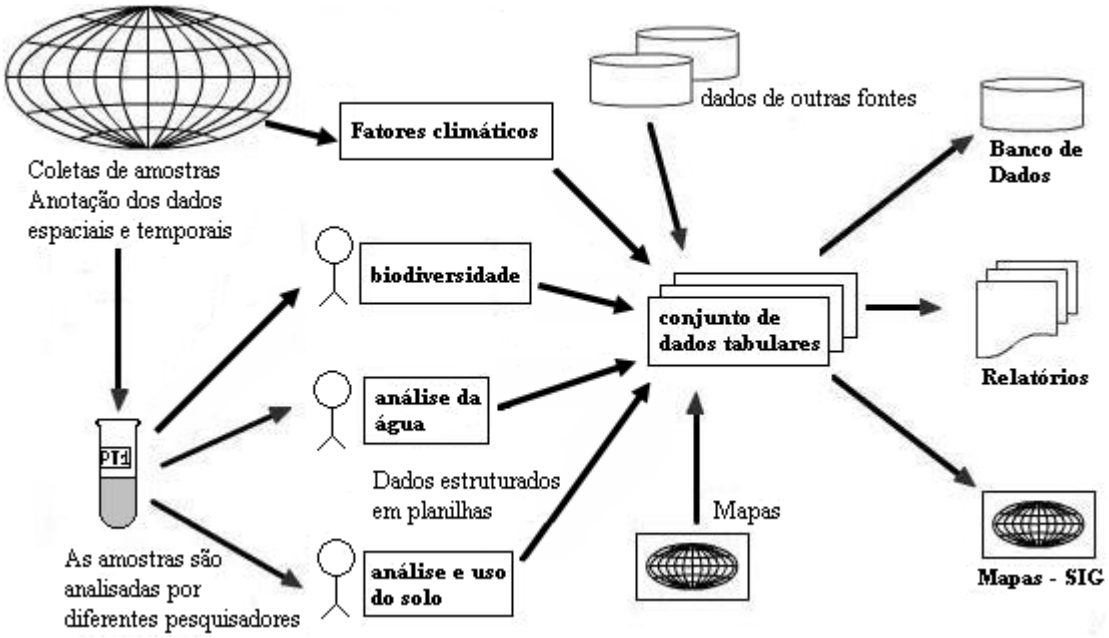


Figura 2.3 – Um exemplo típico de como os dados do domínio são trabalhados.

O fluxo da Figura 2.3 segue a seguinte descrição: primeiramente os dados são obtidos. A obtenção dos dados depende de seu tipo, podendo iniciar-se em coletas de amostras em campo, seguindo das análises para obter os valores dos parâmetros ou são obtidos através de pesquisas em fontes externas que já contenham os valores dos parâmetros, ou ainda através de mapas apropriados ao estudo em questão. Esses dados são estruturados e armazenados geralmente em diversas planilhas eletrônicas, formando um conjunto de dados tabulares. Em alguns casos sua estruturação e armazenamento podem ser transferidos a SGBDs (Sistemas Gerenciadores de Banco de Dados), como mostrado na figura, ou serem diretamente estruturados e armazenados em SGBDs. Os dados são utilizados principalmente na elaboração de relatórios a respeito do estudo realizado, podendo também ser utilizados na construção de mapas e trabalhos em SIGs, provendo assim uma visualização geográfica dos dados analisados.

Nesse exemplo, tornam-se aparentes os principais empecilhos da manipulação dos dados e integração dos mesmos, sendo:

- Estruturação e armazenamento dos dados: poucos trabalhos realizados nesse domínio, com referência nas instituições IIE/IIEGA, têm seus dados estruturados e armazenados em um banco de dados, como em um SGBD. Grande parte dos dados está em planilhas eletrônicas ou mesmo em documentos de texto. Esses dois tipos de fontes de informação são considerados não estruturados, sem uma especificação formal que facilite os relacionamentos entre seus elementos. Tal fato dificulta as pesquisas nessas fontes de dados, principalmente aquelas mais complexas que envolvem relações entre os dados, especialmente nos documentos de textos;
- Distribuição dos dados: uma vez que o trabalho é dividido entre vários agentes, que acabam concentrando individualmente parte dos dados. Assim, os dados, muitas vezes pertinentes a uma única pesquisa, ficam distribuídos entre vários responsáveis e em diferentes locais de armazenamento;
- Heterogeneidade lógica: diz respeito às diferentes estruturas tabulares (quando armazenados em planilhas ou banco de dados) e diferenças nas semânticas em relação aos dados, como homônimos e sinônimos (ex.: dbo é o mesmo que demanda bioquímica de oxigênio), diferentes unidades de medidas (ex.: em uma análise o comprimento foi representado em metros e em outra em quilômetros) e diferenças na contextualização temporal (ex.: a medida de coliformes fecais em duas análises não se alterou, entretanto deve-se observar que são amostras de datas diferentes). Esse obstáculo da heterogeneidade entre as fontes se torna uma motivação para debates posteriores neste documento, relacionados à integração dos dados;
- Grande quantidade de parâmetros e relacionamentos: muitos trabalhos envolvendo esse domínio implicam grandes quantidades de parâmetros e/ou dados sobre tais parâmetros. Como dito no início da seção 2.2, é desejável em um estudo sobre bacia hidrográfica o maior número possível de dados, uma vez que estes números estão relacionados à qualidade do estudo em questão. Além disso, muitas vezes o estudo sobre bacias hidrográficas envolve grandes áreas geográficas, o que é um possível fator de aumento na quantidade de dados necessários. Também nessa questão está a complexidade das relações que envolvem o domínio de bacias hidrográficas, uma vez que ela está relacionada com todo o ambiente que a compõe. Essa complexidade dificulta a modelagem das bacias hidrográficas em um estudo mais completo.

2.4. Conclusão do capítulo

Neste capítulo foi apresentado o domínio do caso de estudo, de forma geral, abstraindo-o, definindo apenas alguns pontos que são mais importantes no desenvolvimento do estudo. Contudo, os conhecimentos na área da ciência da computação apresentados neste documento pretendem ser de utilidade mais geral e não fechada nesse domínio específico.

Uma grande quantidade de parâmetros envolve os estudos relacionados a esse domínio. Entre estes dados existem atributos para a localização espacial das instâncias. Essa característica os distingue de outros tipos de dados. Na área da ciência da computação esse tipo de dado é geralmente chamado de dado espacial. Uma outra característica presente é a existência de atributos para a identificação das instâncias no tempo. Isso proporciona uma análise histórica dos dados, além de ser necessária para relacionar diferentes parâmetros.

Dificuldades em se trabalhar com os dados desse domínio são constantemente enfrentadas. Relacionadas a fatores como estruturação, armazenamento, distribuição, heterogeneidades e grande quantidade dos dados. Para enfrentar essas dificuldades uma modelagem do domínio, que represente seus parâmetros e relações de forma clara, para o entendimento tanto pelo humano quanto pela máquina, é extremamente útil senão necessária. Isso está sendo conseguido com o emprego de ontologias nas áreas da ciência da computação e afins, sendo esse o tópico do próximo capítulo.

3. Ontologia

O sistema DISFOQuE é baseado em ontologia, sendo esta a provedora da visão do domínio a ser integrado. Assim, qualquer que seja o domínio integrado pelo DISFOQuE é necessário a utilização de uma ontologia. Dessa forma, uma ontologia foi construída para representar o domínio do caso de estudo. Também uma outra ontologia foi desenvolvida para um outro domínio, com intuito de testar a utilização e desempenho do sistema em um ambiente diferente.

Desse modo, este capítulo traz uma narrativa sobre o tema ontologia. A seção 3.1 inicia descrevendo o conceito de ontologia na ciência da computação seguindo de sua constituição, descrevendo seus componentes e os conceitos que ela implementa. Após as definições iniciais, na seção 3.2 é apresentado onde a ontologia pode ser utilizada na computação, criando a motivação do uso da ontologia em sistemas de integração como uma representação do domínio. Posteriormente, na seção 3.3 se discute os tipos de ontologias, classificando-as de acordo com sua abordagem. Após esses conceitos teóricos, é realizada na seção 3.5 uma descrição mais técnica expõe as linguagens formais para escrita de ontologias, detalhando a linguagem OWL, além de descrever a base para criação de ontologias, apresentando a metodologia METHONTOLOGY na seção 3.4. Para finalizar o capítulo, na seção 3.6 são dadas as conclusões necessárias que motivam o uso da ontologia em sistemas de integração, sendo este o foco do trabalho.

3.1. Definições e constituição de ontologia

A definição inicial de ontologia vem da antiga filosofia grega, derivada das palavras ont(o)- + -logia, significando o estudo do conhecimento do ser, tratando da natureza do ser, de sua realidade, da existência dos entes, tendo por objetivo o estudo das propriedades mais gerais do ser. É definida pelo dicionário Aurélio como: “parte da filosofia que trata do ser enquanto ser, i. e., do ser concebido como tendo uma natureza comum que é inerente a todos e a cada um dos seres” [1]. Segundo Guarino e Giaretta [4], nesse caso a palavra Ontologia deve ser escrita com “O” maiúsculo. Esses mesmos autores agrupam, além dessa interpretação da Ontologia como uma disciplina filosófica, outras interpretações comuns ao termo e que se adaptam melhor na definição de ontologia empregada na computação. Portanto, o termo ontologia empregado na área da computação, mesmo não tendo uma

definição propriamente única, é diferente da definição da antiga filosofia grega, ainda que tenha uma concepção comum.

No campo da computação, a definição de ontologia mais citada na literatura é a dada por Thomas R. Gruber (1993) [5], juntamente com a atualização de W. N. Borst (1997) [6] e posteriormente por Studer, Benjamins e Fensel (1998) [7]. Tal definição de Gruber é escrita a seguir:

"Uma ontologia é uma especificação explícita de uma conceitualização".

Em 1997 W. N. Borst, na sua tese [6] define a ontologia, baseado na definição de Gruber, como sendo:

"Uma ontologia é uma especificação formal de uma conceitualização compartilhada".

Depois, em 1998, no artigo [7], Rudi Studer, V. Richard Benjamins e Dieter Fensel, baseados nessas duas definições anteriores de Gruber e Borst, definiram ontologia como:

"Uma especificação explícita e formal de uma conceitualização compartilhada".

E essa definição de [7] que será a utilizada neste documento para se referir ao termo ontologia. Nela, a palavra especificação se refere aos conceitos, propriedades, relações, funções, restrições e axiomas definidos; explícitos que estes conceitos e restrições são definidos claramente; formal significa que estas especificações sejam compreendidas e manipuladas por máquina; conceitualização diz respeito a um modelo abstrato de algum fenômeno do mundo real e compartilhado refere-se ao conhecimento consensual em uma comunidade [7].

Definido o termo ontologia, chegando a uma interpretação do termo que será usada neste documento, é necessário expor quais os componentes de uma ontologia. Dessa forma, se concretiza uma idéia mais sólida do que é e o que faz a ontologia. Basicamente uma ontologia é composta por classes, propriedades, restrições e instâncias [8], detalhadas abaixo:

- Classes (conceitos): muitas vezes são o foco das ontologias, descrevendo os conceitos do domínio. As classes podem possuir uma classificação hierárquica, com superclasses e subclasses, de forma consecutiva;
- Propriedades (*slots*): são as características que descrevem cada conceito. Podem ser características que se relacionam apenas com o próprio conceito (atributos) ou que relacionam conceitos diferentes (relações);
- Restrições (facetas ou axiomas): impõem condições sobre os conceitos e propriedades, de forma que a máquina possa interpretá-los automaticamente, garantindo a integridade das instâncias [9];

- Instâncias (indivíduos): representam os elementos de uma ontologia, ou seja, são as ocorrências dos conceitos e propriedades que foram estabelecidas pela ontologia. Em outras palavras, uma instância é um conceito que pertence a uma classe e que possui determinados valores de propriedades.

Para exemplificar a constituição de uma ontologia, a Figura 3.1 apresenta do lado esquerdo uma ontologia abstrata, apenas como didática para representar a ontologia do lado direito relativa à parte de uma modelagem de um domínio do mundo real. Nessa figura, o relacionamento *isa* representa a relação hierárquica entre classes e o relacionamento *io* indica que determinada instância pertence a determinada classe.

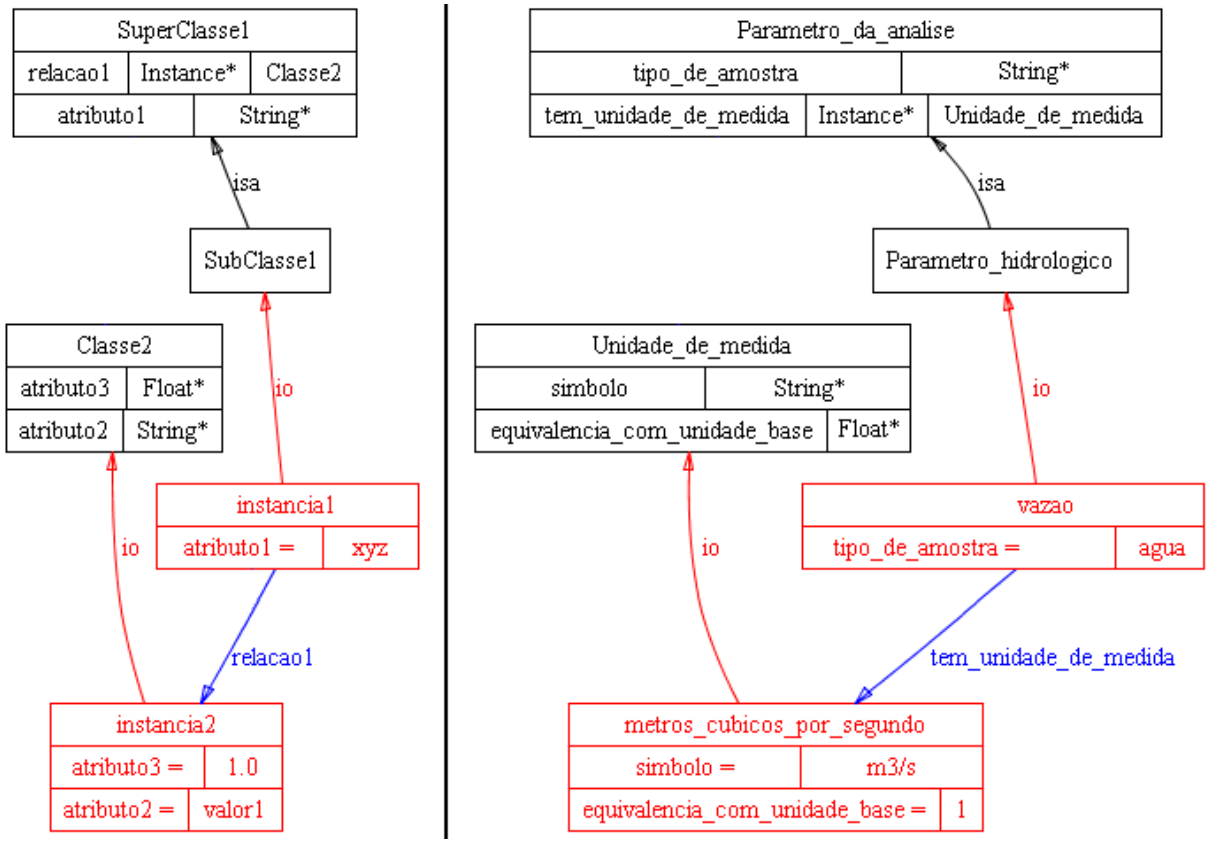


Figura 3.4 – Representações gráficas de componentes de ontologias (figura gerada no software Protégé/Ontoloviz).

Observando-se a constituição de ontologias, destacam-se alguns relacionamentos para representar conhecimentos incluídos na ontologia, sendo estes [10]:

- Taxonomias: representa as hierarquias entre as classes, através do conceito *"é um"* ou *"tipo de"*. É um sistema de classificação que agrupa e organiza o conhecimento em um domínio usando relações de generalização/especialização através de herança simples/múltipla. Por exemplo, *unidade de tempo é uma unidade de medida*;

- Partonomia e Mereologia: a partonomia implementa o conceito de *"parte de"* com suas várias derivações como: *"componente de"*, *"porção de"*, *"membro de"*, *"área de"* etc. Um exemplo de partonomia é visto na relação *latitude é parte de coordenada*. Já a mereologia, consiste em estabelecer uma estrutura completa com entendimento de todas as possíveis relações de *"parte do todo"*, ou seja, é uma partonomia completa. Como no caso onde *o índice de qualidade da água é constituído por todos os nove parâmetros de análise da água utilizados em seu cálculo*;
- Cronológica: estabelece uma relação de precedência (no tempo) entre os conceitos relacionados. Como o conceito de que *coleta de amostra se antecede à análise da coleta*;
- Topologia: estabelece os limites e fronteiras entre os objetos da ontologia, definindo a teoria das conexões com o conceito *"conectado a"*. Por exemplo, relacionando *as bacias hidrográficas conectadas entre si*.

3.2. Uso da ontologia na computação

Várias são as áreas na computação que vêm pesquisando e utilizando ontologias como um meio de auxílio em suas aplicações, questões e estudos; desenvolvendo métodos para solução de problemas. Cada vez mais, ontologia vem sendo assunto de discussão na comunidade científica, vista como uma alternativa para solucionar problemas novos ou já comuns da computação.

Pelo fato da ontologia estar descrevendo o conhecimento, foi inicialmente referenciada na comunidade da Ciência da Computação na área da Inteligência Artificial, formalizando o conhecimento para a máquina. Pelo mesmo fato é empregada na Engenharia de Software como uma rica documentação do domínio, para a aquisição de requisitos e outras etapas da construção de softwares.

Uma grande variedade de sistemas de informação vem utilizando ontologias de formas diversas. A ontologia pode ser empregada tanto no uso do sistema quanto no seu desenvolvimento (guiando a composição ou apoiando a construção do modelo conceitual). Por exemplo, ontologias são empregadas em interfaces cooperativas, que têm como características esconder do usuário detalhes internos do sistema, tornar o processo de busca fácil e transparente e ajudar o usuário, fornecendo-lhe informações complementares, mesmo sem a intervenção do mesmo. Outro exemplo é o emprego da ontologia no sistema FOQuE

(*Fuzzy Ontology-based Query Expansion*) [11], que acrescenta lógica difusa à ontologia para expandir a consulta e posteriormente classificar os resultados, respeitando a semântica dos dados. O sistema FOQuE será comentado mais adiante neste documento, sendo parte dele integrada ao DISFOQuE, fato que dá origem ao nome de tal SID.

A *Web* Semântica [12] [13], idealizada por Tim Berners-Lee, emprega ontologia em sua constituição. Esse fato trouxe destaque ao termo ontologia na comunidade da Ciência da Computação. Na *Web* Semântica a ontologia vem como uma camada de expressividade semântica para as informações na *Web*. Assim, os agentes de software podem buscar as informações pedidas ou, caso necessário, inferir novos fatos, baseando-se nas ontologias, que modelam determinado domínio.

Na área de banco de dados, as ontologias trazem, como já mencionado, uma semântica aos dados armazenados. Sendo os dados definidos como registros capturados (anotações diretas) e as informações como já tendo um significado (o tratamento dos dados, produzindo deduções e inferências lógicas e confiáveis); então as ontologias são responsáveis pela semântica necessária aos dados para que se obtenha informações sobre estes.

Também na área de banco de dados, o uso da ontologia vem sendo explorado para aplicações de integração de dados. A ontologia (uma única ou um conjunto delas) é a base de alguns SIDs, que utilizam-na como a visão única e transparente do domínio. É principalmente nesse contexto que se enquadra a utilização da ontologia neste trabalho. Um maior destaque no emprego de ontologia em SIDs será dado posteriormente neste documento, em especial no seu uso ao tratamento dos casos de heterogeneidades entre as fontes de dados (seção 5.5).

Enfim, a ontologia vem sendo aplicada em projetos de diferentes domínios, tais como: gestão do conhecimento, comércio eletrônico, processamento de linguagens naturais, recuperação de informação na *Web*, projetos educacionais, entre outros [55].

3.3. Tipos de ontologias

Não existe apenas uma proposta para a classificação de ontologias. Algumas propostas classificam as ontologias conforme a sua função ou o grau de formalismo de seu vocabulário ou ainda a sua aplicação. Entretanto é muito comum na literatura a classificação feita por Guarino (1997, 1998) [15] [16], classificando-as conforme o conteúdo da conceitualização, como descrita a seguir:

- Ontologias de nível superior (gerais): classificam as diferentes categorias que existem no mundo. Representam as noções gerais, que são independentes de um domínio ou problema particular. Os conhecimentos representados nesse tipo de ontologia são utilizados por vários domínios. Conceitos de tempo e espaço são exemplos encontrados nesse tipo;
- Ontologias de domínio: são ontologias mais específicas, que representam o conhecimento de um domínio específico. Essas ontologias descrevem o vocabulário relacionado a um domínio genérico, sobre os conceitos, relacionamentos ou teorias que governam tal domínio. Como exemplo desse tipo, pode-se citar uma ontologia que descreve o vocabulário usado para o domínio de bacias hidrográficas, relacionando os termos, definindo os sinônimos etc;
- Ontologias de tarefa: descrevem vocabulário relacionado a uma tarefa genérica, tal como, diagnose ou vendas;
- Ontologias de aplicação: descrevem parte do conhecimento dependente de um domínio e de uma tarefa particular, relacionando-se, então, a um método de solução de problema em um domínio específico. Nesse tipo, poderia ser classificada uma ontologia que descrevesse o processo que um indivíduo deve desempenhar para realizar uma atividade como análise da qualidade da água;
- Ontologias de representação (meta-ontologias): capturam as representações das primitivas, usadas para formalizar conhecimento em um dado sistema ou família de representação do conhecimento. Uma ontologia que define os conceitos da lógica difusa é um exemplo.

Guarino (1998) [15] especifica uma hierarquia entre os tipos de ontologia, segundo a qual os conceitos das ontologias de tarefa e domínio devem ser especializações das ontologias de nível superior e por sua vez os conceitos das ontologias de aplicação, especializações das ontologias de domínio e tarefa. A Figura 3.2 ilustra esta hierarquia.

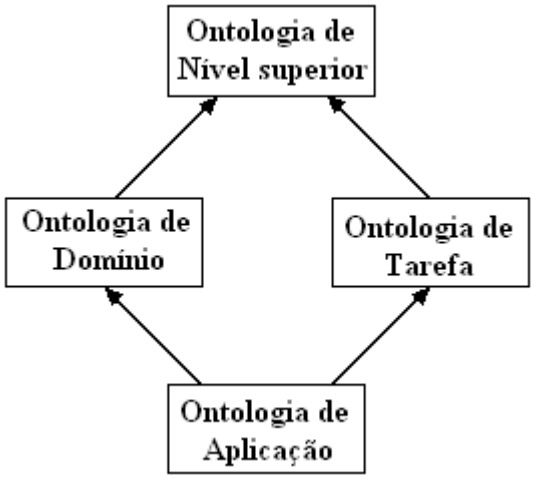


Figura 3.5 – Tipos de ontologias, conforme seu nível de dependência em relação a uma tarefa ou ponto de vista particular. (extraída de [15])

Conhecidos os tipos e características de ontologias, podem ser pesquisadas ontologias existentes adequadas à utilização desejada. No caso deste trabalho de mestrado, foi desenvolvida uma ontologia de domínio, sendo este o tipo mais comum, utilizada em um micro-mundo [17], como o domínio de bacias hidrográficas.

3.4. Metodologias de construção de ontologias

Várias metodologias existem na literatura a respeito da construção de ontologias. Assim como nas metodologias de desenvolvimento de software, muitos dos métodos de desenvolvimento de ontologias requerem um processo interativo, com revisões constantes. Porém, mesmo com vários métodos existentes, o desenvolvimento de ontologias envolve basicamente:

- Determinar o domínio e escopo: sendo o domínio a parte mais ampla, definindo o objeto principal e o escopo especializa, dentro do domínio, qual o objetivo daquela especificação. Como exemplo, o domínio sendo o de bacias hidrográficas e o escopo sendo a avaliação das bacias em relação aos atributos físico-químicos e biológicos e seus relacionamentos;
- Definição das classes: avaliar quais as classes dentro do domínio e a pertinência destas no escopo, dependendo de quão amplo são o domínio e escopo. Exemplos de classes do domínio de bacias hidrográficas com escopo em sua análise são: *rio*, *riacho*, *parâmetro de análise*;
- Organização das classes em uma taxonomia: determinar as classes e subclasses. Assim, como exemplo, a classe *sub-bacia* seria subclasse da classe *bacia*;

- Definição das propriedades: quais os atributos das classes e suas relações. Como exemplo, a criação do atributo *tipo de amostra* para a classe *parâmetro de análise* e a relação *possui unidade* entre as classes *parâmetro de análise* e *unidade de medida*;
- Definição dos axiomas: as sentenças que são sempre uma verdade, como em: *todo peixe é um animal aquático*;
- Definição de funções: os cálculos, como exemplo, a determinação do índice de qualidade da água para consumo humano, envolvendo nove parâmetros da água analisada;
- Definição das instâncias: criação dos elementos. Exemplo, bacia *Paraná*, rio *Tietê*, parâmetro *fósforo total*.

Visto as questões básicas e princípios que envolvem a construção de ontologias, será apresentada uma entre as várias metodologias existentes. A metodologia apresentada será a METHONTOLOGY, criada pelo grupo de Engenharia de Ontologias da Universidade Politécnica de Madri [14]. A ontologia do caso de estudo deste trabalho foi desenvolvida observando-se as questões básicas e baseando-se no método da METHONTOLOGY.

A METHONTOLOGY divide-se em três grupos de atividades, sendo: atividades de gerenciamento, atividades de desenvolvimento e atividades de suporte.

No grupo de atividades de gerenciamento estão as atividades de:

- Planejamento: quais tarefas serão desenvolvidas;
- Controle: garantia de que as tarefas planejadas serão executadas;
- Garantia da qualidade: desenvolver a ontologia como estabelecido.

No grupo de atividades de desenvolvimentos tem-se:

- Especificação: o propósito e usuários finais da ontologia a ser desenvolvida;
- Conceitualização: criação do modelo conceitual;
- Formalização: transformação do modelo conceitual em um modelo formal;
- Implementação: escrita do modelo formal em linguagens computáveis;
- Manutenção: correção e alterações na ontologia vistas como necessárias.

No grupo de atividades de suporte tem-se:

- Aquisição de conhecimento: busca do conhecimento do domínio da ontologia;
- Avaliação: comparações técnicas das ontologias, relacionadas com o software e documentação;
- Integração: uso de ontologias já existentes;
- Documentação: descrição das fases;

- Gerenciamento de configurações: registra as alterações no software, documentação e ontologia.

A Figura 3.3 apresenta essas atividades durante o processo de desenvolvimento cíclico da ontologia.

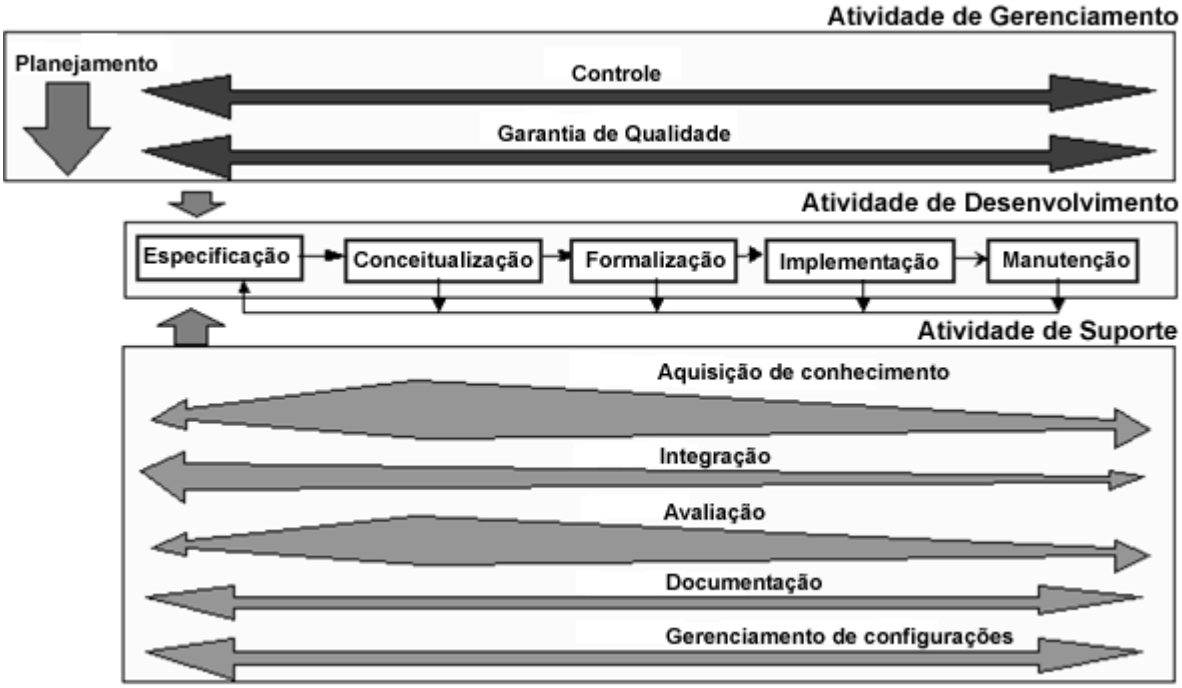


Figura 3.6 – O ciclo de desenvolvimento da ontologia na METHONTOLOGY. (extraída de [14])

Observa-se a semelhança dessa metodologia com processos de desenvolvimento de software como o processo RUP (análise, projeto, implementação, teste com concepção, elaboração, construção, transição).

Outras metodologias como Enterprise Ontology, TOVE, On-To-Knowledge entre outras, são encontradas na literatura sobre o assunto, sendo que cada grupo aplica sua própria abordagem, sem ainda uma proposta unificada.

No tópico seguinte será discutido como implementar de maneira formal as ontologias (linguagens de implementação).

3.5. Linguagens de representação de ontologias

Para que as ontologias sejam compreendidas pelas máquinas, respeitando o formalismo descrito em sua definição, devem ser escritas em linguagens formais, que se atêm a fórmulas claramente estabelecidas. Respeitando-se esse princípio, várias linguagens para implementar as ontologias são propostas, provendo diferentes facilidades. Grande parte das

linguagens é baseada na lógica de descrição ou extensões dela, tendo como exemplos desse grupo de linguagens a LOOM e CARIN. Um outro importante grupo de linguagens são as baseadas em *frames* (quadros, "classes"), as *frame-based*, tendo como exemplo a linguagem F-logic. Outras linguagens juntam as primitivas, típicas das linguagens *frame-based* e a capacidade de inferência da lógica de descrição, como é o caso da linguagem OIL (*Ontology Inference Layer*) [55].

A lógica de descrição é baseada nos predicados. Os predicados unários representam as classes e as propriedades, sendo atributos das classes. Os predicados binários relacionam duas classes. Já os *frames* representam as classes e cada classe tem seu conjunto de atributos, que estão restritos a ela. Assim, esses sistemas *frame-based* representam objetos complexos do mundo real, no entanto não tratam adequadamente conceitos, como exceção e inconsistência [18].

Dentre as linguagens de implementação de ontologia, a linguagem OWL (*Web Ontology Language*) é das que mais se destacam na literatura. Isso se deve ao fato dessa linguagem ser a proposta de padrão do grupo W3C (*World Wide Web Consortium*) como a linguagem para a *Web* Semântica. Ela deriva-se de duas propostas anteriores, a linguagem OIL e a DAML (DARPA *Agent Markup Language*).

A OIL foi a primeira delas, que como já mencionado, permite definições baseadas em *frames* implementando as primitivas definidas pelo modelo OKBC, além das definições em lógica de descrição. As definições de ontologias são geradas sobre XML e RDF. A linguagem OIL tem um conceito de modularidade, ou seja, é dividida em camadas de acordo com a complexidade que se deseja de seu uso, garantindo o uso mais simples pelos usuários. Essa linguagem possui um motor de inferência consistente, completo e eficiente, manipulando tanto *frames* como lógica de descrição, no entanto perdeu-se um pouco da expressividade.

A DAML foi um investimento do DARPA (Departamento de Defesa dos Estados Unidos) sobre a OIL, sem promover um motor de inferência. Posteriormente foi criada a linguagem DAML+OIL e depois acrescentado requisitos de internacionalização (Unicode), de apresentação e documentação. Daí se originou a linguagem OWL.

Baseado na linguagem OIL, a OWL também está estruturada sobre XML, XML *Schema*, RDF e RDF *Schema*. O XML provê a sintaxe para documentos estruturados, mas sem imposição de restrições semânticas no significado dos documentos. O XML *Schema* restringe a estrutura de um documento XML. Aparece então o RDF, um modelo de dados para descrever recursos e as relações entre eles, provendo uma semântica simples. O RDF é

baseado na sintaxe XML. Logo surge o RDF *Schema*, um vocabulário para permitir a descrição de propriedades e classes com alguma expressividade limitada. Finalmente, o OWL acrescenta vocabulário e definições mais formais para a descrição de ontologias, como relações entre classes, cardinalidade, igualdade, tipos de propriedades complexas etc [19].

A OWL, assim como a OIL, tem o conceito de modularidade, sendo subdividida em três linguagens:

- OWL *Lite*: é uma sub-linguagem da OWL DL, que usa somente algumas características da linguagem OWL e possui mais limitações do que OWL DL ou OWL *Full*. É específica para necessidades básicas dos usuários, com restrições simples. Cardinalidade é suportada apenas com valores 0 ou 1. A OWL *Lite* é o mais simples a ser implementado;
- OWL DL: permite o máximo de expressividade até o ponto em que é garantida a computabilidade (chega-se a conclusões e em tempo finito). Mas para isso, apesar de permitir todas as construções da linguagem OWL, estas estão sujeitas a certas restrições. A sigla DL possui correspondência com a lógica de descrição (*description logics*);
- OWL *Full*: permite o máximo de expressividade, porém, sem garantia computacional. A OWL *Full* e a OWL DL suportam o mesmo conjunto de construções da linguagem OWL, embora com restrições um pouco diferentes, como exemplo, em OWL DL a cardinalidade não pode ser usada em propriedades transitivas. A falta dessas restrições na OWL *Full* a deixa com problemas de consistência lógica.

Enfim, a OWL define formalmente um conjunto de termos que são usados para representar um domínio específico. Dessa forma, OWL pode ser usada por ferramentas automatizadas, fazendo inferências que melhoraram, por exemplo, a busca de informação e o gerenciamento de conhecimento. É uma linguagem amplamente difundida, facilita o reuso de ontologias e a criação das mesmas por ferramentas, como o Protégé [20]. Com essas características é uma linguagem adequada e utilizada no desenvolvimento das ontologias neste trabalho.

3.6. Conclusão do capítulo

A definição de ontologia para a computação, apresentada como uma especificação explícita e formal de uma conceitualização compartilhada, tem por objetivo e

concepção geral definir os termos usados para descrever e representar um domínio. Ou seja, uma descrição dos conceitos e relacionamentos entre eles, que pode ser usada por pessoas ou agentes de software para compreender e compartilhar informações dentro desse domínio. Centrado nesse conceito, uma ontologia, independente de seu tipo, é constituída por componentes bem definidos, escrita em linguagem formal e possui metodologias para seu desenvolvimento. De fato uma ontologia não é algo trivial a ser construído. É necessário um consenso de uma comunidade apoiada por especialistas do domínio estruturado.

A ontologia vem sendo estudada pela comunidade da computação e, conseqüentemente, amadurecendo em relação a todas as questões abordadas. Entre esses estudos, está a questão do uso da ontologia como base para integração de fontes de dados. É este o foco deste trabalho, utilizar ontologia como base para integração de dados. Assim, o capítulo seguinte é relativo ao tema de integração de dados e no uso de ontologia nos sistemas de integração.

4. Integração de dados

Integração de dados é tema chave do trabalho aqui descrito. O objetivo principal é a integração de dados, em especial de dados referentes à análise de bacias hidrográficas, baseando-se em ontologia para representar o domínio. Para atingir esse objetivo é desenvolvido um Sistema de Integração de Dados (SID). Mas antes de descrever o SID, um estudo sobre o tema de integração de dados e posteriormente sobre sistemas de integração baseados em ontologia é realizado neste capítulo.

Com intuito de servir como base teórica para o desenvolvimento de um SID baseado em ontologia, este capítulo aborda a questão de integração de dados, apresentando inicialmente na seção 4.1 a motivação e a definição para integração de dados. Em seguida na seção 4.2 comenta os principais desafios à integração de dados e na seção 4.3 analisa as principais abordagens para integração de dados. Após descrita essa visão geral sobre especificamente o tema integração de dados, a seção 4.4 inicia o estudo sobre o uso da ontologia em sistemas de integração de dados, qual o interesse e como é realizada esta união entre ontologia e sistemas de integração.

4.1. Motivação e definição

Muitas vezes é desejável, se não necessária, a busca de informações em bases de dados distribuídas e heterogêneas. Essa necessidade é visível no domínio de bacias hidrográficas, especialmente tratado neste trabalho. Porém, não é somente nesse domínio que se deseja uma busca de dados em diferentes fontes. Nos casos que envolvem informações científicas torna-se especialmente proveitosa uma integração de dados, uma vez que há uma quantidade considerável de dados importantes para as análises científicas.

Dessa forma, Sistemas de Integração de Dados são desenvolvidos para prover ao usuário uma visão única, uniforme e homogênea de fontes de dados heterogêneas, desenvolvidas independentemente. O objetivo maior desses sistemas é possibilitar ao usuário acesso às múltiplas fontes de forma transparente, sem que o usuário tenha que localizá-las, interagir isoladamente com cada uma delas e então ter que combinar manualmente os dados encontrados nas fontes [21] [22].

No domínio de análises de bacias hidrográficas um SID seria de grande auxílio na busca de informações para tais análises, sendo o domínio multidisciplinar, uma vez que

trabalha com múltiplas áreas de estudo como: estudo do tipo, uso e ocupação do solo; estudo dos parâmetros avaliadores da qualidade da água; estudo meteorológico; estudo das espécies de peixes etc. No entanto, os dados desses estudos geralmente estão distribuídos em fontes de dados heterogêneas, o que dificulta uma busca isolada por cada um dos dados e o relacionamento entre eles para avaliações históricas ou mesmo em períodos mais recentes.

Assim, os SIDs são de grande apoio em pesquisas que necessitam de dados contidos em múltiplas fontes. Contudo, alguns problemas, como a distribuição e a heterogeneidade, devem ser tratados nesses sistemas. Problemas estes definidos na seção seguinte.

4.2. Desafios da integração de dados

Voltando à definição, integração de dados deve prover ao usuário uma visão única, uniforme e homogênea de fontes de dados heterogêneas, desenvolvidas independentemente. Este ambiente de fontes independentes, ou seja, desenvolvidas de maneira própria sem a preocupação de relacionar-se a outras fontes, possui características que se tornam desafios no momento em que se queira integrá-las. Esses desafios são: a distribuição, a autonomia, a flexibilidade e a heterogeneidade relacionadas às fontes de dados, explicadas nas subseções seguintes.

4.2.1. Distribuição

As fontes de dados se encontram distribuídas, ou seja, em locais fisicamente diferentes e possivelmente distantes entre si. Essa questão de interoperar as fontes dispersas, de forma a compartilhar seus dados, é uma discussão relativa à área de redes de computadores.

É cada vez mais comum o fato dos computadores não estarem trabalhando isoladamente, mas sim ligados em redes, cada vez mais velozes, que provêem a comunicação entre eles. A maioria dos computadores hoje se encontra ligado a alguma rede, destacando-se a Internet (rede mundial de computadores). Assim, essa barreira da distribuição das fontes de dados está sendo suprimida com a evolução das redes de computadores. Essa evolução das redes diz respeito à, principalmente, velocidade, custo de transmissão e disponibilidade.

4.2.2. Autonomia e gerenciamento

Autonomia é o grau de independência dos sistemas de informação que gerenciam as fontes de dados a serem integradas. Medida por fatores como: a possibilidade das fontes trocarem ou não informação, de poderem executar as transações de forma independente e de ter ou não permissão de modificar informações. Özsu e Valduriez (1999) [23] classificaram o grau de autonomia dos sistemas de informação integrados dividindo-os em três classes:

- Sistemas fortemente integrados: é disponível para o usuário uma única visão de toda a base de dados. Entretanto, múltiplas bases de dados podem armazenar as informações dessa visão única. O controle do processamento de cada solicitação do usuário é responsabilidade de um único SGBD. Nesses sistemas fortemente integrados os SGBDs não operam de maneira independente, mesmo tendo capacidade para tal;
- Sistemas semi-autônomos: nesses sistemas os SGBDs podem operar independentemente, porém, ainda participam de uma coleção de sistemas de bancos de dados que cooperam entre si, permitindo o compartilhamento de seus dados, podendo assim ser chamados de bancos de dados federados. Esses bancos possivelmente são heterogêneos. Esses SGBDs têm autonomia para determinar quais partes de seu banco de dados ficará acessível para os usuários de outros SGBDs. Eles não são sistemas totalmente autônomos, pois precisam ser modificados para que possam trocar informações uns com os outros;
- Sistemas totalmente isolados: os componentes individuais são SGBDs *stand-alone* (dedicado, autônomo) que desconhecem a existência de outros SGBDs, bem como a maneira de se comunicar com eles. Nesses sistemas, o processamento de transações do usuário que acessam vários bancos de dados é difícil, pois não existe nenhum controle global sobre a execução de SGBDs individuais.

A autonomia é um fator influente na interoperabilidade entre as fontes de dados, sendo uma barreira maior ou menor dependendo da proporção (diretamente proporcional) do grau de autonomia que possuem os sistemas.

Ao se falar sobre autonomia (seja qual for o nível) de um sistema de informação, infere-se que a base de dados é gerenciada, geralmente por um SGBD. Nesse caso a base de dados é estruturada (modelo relacional ou orientado a objetos) ou pelo menos semi-estruturada (XML). Esse gerenciamento fornece vantagens, como na busca dos dados provendo uma linguagem de pesquisa bem definida (SQL, OQL, XQuery etc). Em fontes de

dados que não possuem um gerenciamento, sendo elas não estruturadas (planilhas, documentos de texto), fica mais difícil o manejo dos dados.

4.2.3. Flexibilidade

É a capacidade do sistema ser flexível, adaptando-se rapidamente a novas fontes de informação e a modificações das fontes existentes. Essa capacidade torna-se uma característica importante, principalmente com o rápido crescimento de fontes de dados disponíveis atualmente.

Assim, deve-se almejar que as soluções que usam integração de dados sejam projetadas com a facilidade de inclusão de novas fontes de dados, já que possivelmente com o avanço das redes, o crescimento de informações disponíveis e a necessidade de comunicação e compartilhamento dos dados, surja o interesse em adicionar novas fontes de pesquisa.

Em relação às alterações das fontes existentes, deve-se ter cautela, uma vez que causam possíveis ajustes no sistema de integração. Entretanto, esse caso de alterações das fontes existentes é mais raro, pois as fontes tendem a ser estruturalmente estáticas, já que geralmente passam por fases de modelagem e revisões antes de serem propriamente criadas.

4.2.4. Heterogeneidade

Por definição, o termo heterogêneo é: “de diferente natureza.” [1]. Nas diferentes fontes de dados de um sistema de informação, a heterogeneidade é causada devido às diferentes soluções das equipes de desenvolvimento de cada fonte. Isso é natural, pois cada equipe tem sua forma de abstrair o domínio, tendo requisitos específicos, modelagens diferentes e mesmo tecnologias de implementação diversas. Entretanto, esse é geralmente o principal desafio a ser resolvido pelos Sistemas de Integração de Dados.

Segundo Busse (1999) [24], a heterogeneidade dos sistemas de informação é classificada em:

- Heterogeneidade sintática: aborda problemas sintáticos relacionados a aspectos técnicos, os quais são divididos em dois:
    - Heterogeneidade técnica: trata do nível de sistema com: diferentes plataformas de hardware (SPARC, desktop etc), sistemas operacionais (Unix, Windows, Linux etc), protocolos (http, ODBC, entre outros) e linguagens de programação (Java, C, C++ etc);

  - Heterogeneidade de interface de acesso: trata das diferenças no acesso aos componentes, como as diferentes linguagens de consulta (SQL, OQL e demais) e as restrições diferentes à consulta, como operações não permitidas a determinados bancos.
- Heterogeneidade de modelo de dado: trata do fato do uso de diferentes modelos de dados, como o modelo relacional e orientado a objetos;
- Heterogeneidade lógica: ocorridas no nível de esquema ou de dados, divididas em:
  - Heterogeneidade semântica: diferenças nas semânticas em relação aos dados, com respeito aos significados, interpretações ou uso pretendido do dado;
  - Heterogeneidade estrutural: existe quando elementos do mundo real são modelados com o mesmo modelo de dados, mas com diferentes conceitos de modelagem;
  - Heterogeneidade esquemática: é um tipo especial de heterogeneidade estrutural. Esse conflito ocorre quando os conceitos são modelados usando diferentes elementos de um modelo de dados. Como no caso do modelo relacional, no qual em uma base de dados o elemento do mundo real é modelado como um atributo de uma entidade e em outra é modelado como valores de um atributo de uma entidade. Esse tipo de heterogeneidade é tratado separadamente, pois causa conflito entre dados e esquema.

A heterogeneidade lógica causa conflitos que dificultam a integração dos dados. Em um artigo dos autores Sheth e Kashyap (1992) [26] é feita uma enumeração e classificação dos conflitos lógicos (semântico, estrutural e esquemático), provenientes das heterogeneidades entre as fontes. A Tabela 4.1 apresenta a classificação, entre tipos de incompatibilidade, dos conflitos causados pela heterogeneidade apresentados por Sheth e Kashyap [26], juntamente com o tipo de heterogeneidade lógica apresentado por Busse em [24].

Tipo de Heterogeneidade Lógica	Tipo de Incompatibilidade	Conflitos
Heterogeneidade semântica	Incompatibilidade de definição de domínio	Conflitos de nome
		Conflitos de representação de dados
		Conflitos de escala
		Conflitos de precisão
		Conflitos de valor padrão
		Conflitos de restrição
	Incompatibilidade de valores de dados	Conflitos de inconsistência conhecida
		Conflitos de inconsistência temporal
		Conflitos de inconsistência aceitável
Heterogeneidade estrutural	Incompatibilidade de definição de entidade	Conflitos de identificadores
		Conflitos de nome
		Conflitos de compatibilidade de união
		Conflitos de isomorfismo de esquemas
		Conflitos de ausência de dados
	Incompatibilidade de nível de abstração	Conflitos de generalização
		Conflitos de agregação
Heterogeneidade esquemática	Discrepância esquemática	Conflitos de valor e atributo
		Conflitos de atributo e entidade
		Conflitos de valor e entidade

Tabela 4.1 – Conflitos causados pelas heterogeneidades lógicas.

Devido à relevância dos conflitos causados pela heterogeneidade lógica para o trabalho aqui apresentado, uma explicação de cada um dos conflitos apresentados na Tabela 4.1 será feita a seguir. Como motivação para futuras seções neste documento, os exemplos utilizados são do domínio do caso de estudo de análise de bacias hidrográficas.

- Conflitos de nome: ocorrem devido aos homônimos e sinônimos. Inclui-se nesse tipo de conflito a nomeação de entidades, atributos e valores.

  Exemplo: Duas bases de dados com as entidades *tb_corpo_dagua* e *rio*, sendo elas semanticamente semelhantes, o mesmo ocorre com os atributos das bases de dados, como *tipo* e *classe* e também com os valores dos dados, como em *Rio Tietê* e *tiete*, sendo os dois o mesmo rio.

Base de Dados A
tb_corpo_dagua

ID	nome	tipo	bh
1	Rio Tietê	Rio	Paraná

Base de Dados B
rio

ident	nome	classe	bacia
5	tiete	rio	parana

Figura 4.7 – Conflitos de nomes.

- Conflitos de representação de dados: ocorrem com os tipos de dados diferentes, como em um atributo *latitude* sendo definido como do tipo *float* em uma base de dados e em outra sendo definido como do tipo *string*.
- Conflitos de escala: ocorrem devido ao emprego de valores de unidades diferentes entre os atributos semanticamente iguais.

  Exemplo: Duas bases de dados com o atributo *comprimento*, entretanto em uma das bases de dados os valores do atributo é dados em metros e na outra em quilômetros.

Base de Dados A
rio

ID	nome	comprimento
1	Tiete	1000

Base de Dados B
rio

ID	nome	comprimento
1	Tiete	1

Figura 4.8 – Conflitos de escala.

- Conflitos de precisão: ocorrem quando um valor representa uma faixa de valores.

  Exemplo: Duas bases de dados, ambas com o atributo *iqa*, que indica o Índice de Qualidade da Água. Entretanto, em uma base o *iqa* é dado por seu valor numérico calculado e na outra este atributo é dado pelo estado em que a água é classificada por tal índice.

Base de Dados A
iqa_ponto

ID	ponto	iqa
1	Pt01	90
2	Pt02	30

Base de Dados B
iqa_ponto

ID	ponto	iqa
1	Pt01	Ótima
2	Pt02	Ruim

Relação entre valor numérico e classificação do IQA

Faixa	Classificação
79 < IQA ≤ 100	Ótima
51 < IQA ≤ 79	Boa
36 < IQA ≤ 51	Regular
19 < IQA ≤ 36	Ruim
IQA ≤ 19	Péssima

Figura 4.9 – Conflitos de precisão.

- Conflitos de valor padrão: ocorrem com a definição de diferentes valores padrões (*default*) entre atributos semanticamente semelhantes. Como por exemplo, em uma base de dados o valor padrão para o atributo *altitude* é de *0 metros*, enquanto que em outra base o valor padrão para o mesmo atributo é de *800 metros*.
- Conflitos de restrição: quando bases de dados estabelecem restrições conflitantes a atributos semanticamente semelhantes. Como exemplo, restringir a aceitar apenas os dados de coleta com profundidade acima de 100 metros em uma base de dados e na outra restringir esse valor para dados abaixo dos 90 metros.

- Conflitos de inconsistência conhecida: valores em uma base de dados sabidamente errôneos, confiando-se em outra base de dados. Como no caso em que duas bases de dados armazenam os valores de temperatura do ar, entretanto os valores de dias iguais estão discrepantes em um alto grau. Porém, sabe-se qual a base de dados contém os valores corretos de temperatura do ar.
- Conflitos de inconsistência temporária: quando os valores em uma base de dados então inconsistentes por ainda não estarem atualizados. Por exemplo, a quantidade de peixes coletados de determinada espécie estar inconsistente por ainda não ter sido atualizada com os novos dados da coleta mais recente.
- Conflitos de inconsistência aceitável: ocorrem quando os valores inconsistentes de uma base de dados estão dentro de um limite aceitável. Como no caso em que os dados de temperatura do ar estão inconsistentes em alguns microgrаus Celsius, aceitáveis para o tipo de análise que se quer realizar.
- Conflitos de identificadores: ocorrem na utilização de atributos identificadores (chaves) diferentes para identificar atributos semelhantes.

  Exemplo: Duas bases de dados com a mesma entidade *ponto_coleta*, sendo que em uma das bases a chave primária da entidade é o atributo *ID* gerado automaticamente a cada inserção de dados e na outra base de dados a chave primária é formada pelos atributos *latitude* e *longitude* que identificam unicamente um ponto de coleta.

Base de Dados A

ponto_coleta

ID	latitude	longitude	altitude
1	25:35:04.2	47:05:55.2	25

Base de Dados B

ponto_coleta

latitude	longitude	altitude
253504.2	470555.2	25

Figura 4.10 – Conflitos de identificadores (chaves).

- Conflitos de compatibilidade de união: ocorrem quando os atributos de uma entidade não são semanticamente relacionados com os atributos da outra entidade. Porém, podem existir atributos semanticamente relacionados entre as duas entidades.

  Exemplo:

  Base de Dados A: {PARAMETRO_ANALISE (ID, pH, DBO, Fósforo)}

  Base de Dados B: {PARAMETRO_ANALISE (ID, pH, DBO, Nitrogênio)}
- Conflitos de isomorfismo de esquemas: ocorrem entre entidades semanticamente similares, mas com quantidades de atributos diferentes.

Exemplo: Duas bases de dados com a mesma entidade *comprimento_peixe*, sendo que em uma há apenas o atributo *comprimento_total* para armazenar o comprimento do peixe e na outra base existem os atributos *comprimento_corpo*, *comprimento_cabeca* e *comprimento_cauda*.

Base de Dados A

comprimento_peixe

ID	especie	Comprimento_total
1	5	50

Base de Dados B

comprimento_peixe

ID	especie	comprimento_cabeca	comprimento_corpo	comprimento_cauda
1	5	10	30	10

Figura 4.11 – Conflitos de isomorfismo de esquemas.

- Conflitos de ausência de dados: ocorrem entre duas entidades semanticamente semelhantes quando em uma delas há a ausência de um ou mais atributos. Um caso especial desse conflito é quando se torna possível inferir o valor do atributo faltante através da semântica da entidade.

  Exemplo: Duas bases de dados que possuem entidades semelhantes com dados sobre análises feitas em rios. Entretanto, uma das entidades não possui o atributo que identifica em qual o rio foi realizada a análise. Porém, nessa entidade todos os dados são referentes ao rio Tietê, ou seja, infere-se que o valor para o atributo *rio* ausente é *Tietê* em todas as instâncias.

Base de Dados A

analise

ID	ID_ponto	rio	pH	fosforo
1	5	Tietê	2.4	0.001
2	3	Jacaré	2.1	0.02

Base de Dados B

analise_rio_tiete

ID	ID_ponto	pH	fosforo
1	7	3.2	0.03
2	10	2.7	0.01

Figura 4.12 – Conflitos de ausência de dados.

- Conflitos de generalização: quando é feita uma generalização na abstração de uma entidade em relação à mesma entidade em outra base de dados.

  Exemplo:

  Base de Dados A: {PT_COLETA (ID, Sigla, Latitude, Longitude)}

  Base de Dados B: {PT_COLETA (ID, Sigla, Latitude, Longitude, Altitude)}

- Conflitos de agregação: ocorrem quando em uma base de dados há uma entidade que representa o conjunto de uma outra entidade em outra base de dados.

Exemplo: Uma base de dados com a entidade *peixe* que contém atributos relacionados a cada peixe em particular analisado. Uma outra base de dados com a entidade *peixes*, contendo atributos relacionados a um conjunto de peixes de cada espécie analisada.

Base de Dados A: {PEIXE (ID, espécie, comprimento_total)}

Base de Dados B: {PEIXES (ID, espécie, quantidade, comprimento_medio)}

- Conflitos de valor e atributo: ocorrem quando a semântica de um mesmo objeto do mundo real é representada em uma base de dados por atributos e em outra por valores de atributo.

  Exemplo: Duas bases de dados, ambas com dados sobre o uso de solo. Entretanto uma representa o tipo de uso do solo como sendo valores de um atributo de uma entidade e na outra base como atributos de uma entidade.

Base de Dados A

uso_do_solo

ID	tipo_uso	porcentagem
1	plantio	40
2	pastagem	60
3	plantio	30
4	pastagem	70

Base de Dados B

uso_do_solo

ID	plantio	pastagem
1	40	60
2	30	70

Figura 4.13 – Conflitos de valor e atributo.

- Conflitos de atributo e entidade: semelhante ao anterior, porém em uma base a semântica do objeto é representada por atributos e na outra por entidades.

  Exemplo: Duas bases de dados, ambas com dados sobre o uso de solo. Entretanto uma representa o tipo de uso do solo como sendo atributos de uma entidade e a outra base como sendo entidades distintas.

Base de Dados B

uso_do_solo

ID	plantio	pastagem
1	40	60
2	30	70

Base de Dados C

plantio

ID	porcentagem
1	40
2	30

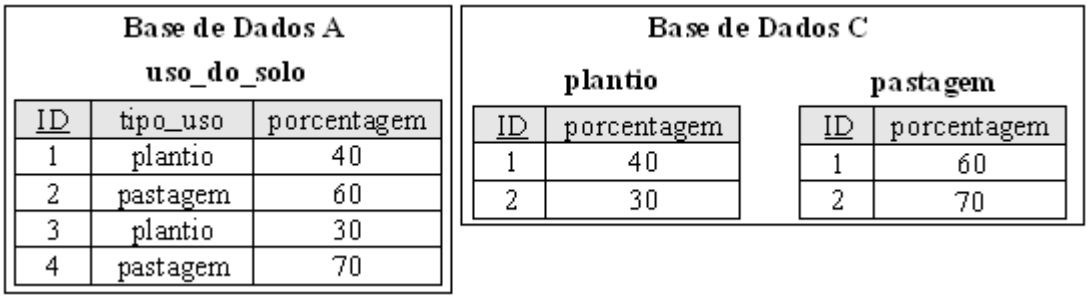
pastagem

ID	porcentagem
1	60
2	70

Figura 4.14 – Conflitos de atributo e entidade.

- Conflitos de valor e entidade: semelhante aos dois anteriores, entretanto a semântica do objeto é representada por valores de atributo em uma base e entidades na outra.

  Exemplo: Duas bases de dados, ambas com dados sobre o uso de solo. Entretanto uma representa o tipo de uso do solo como sendo valores de um atributo de uma entidade e a outra base como sendo entidades distintas.



Base de Dados A

uso_do_solo

ID	tipo_uso	porcentagem
1	plantio	40
2	pastagem	60
3	plantio	30
4	pastagem	70

Base de Dados C

plantio

ID	porcentagem
1	40
2	30

pastagem

ID	porcentagem
1	60
2	70

Figura 4.15 – Conflitos de valor e entidade.

Todos esses casos de heterogeneidade são encontrados nas fontes do domínio estudado. Além do desafio da heterogeneidade, os outros desafios descritos nas subseções anteriores também são obstáculos enfrentados pelo DISFOQuE.

4.3. Abordagens para integração de dados

Basicamente existem dois tipos de abordagens adotadas por um Sistema de Integração de Dados: a virtual e a materializada. Uma outra abordagem, adotada em sistemas mais complexos, implementa ambas as abordagens, chamando-se assim de abordagem híbrida. Essa divisão entre as abordagens é feita observando-se de onde se extrai os dados no momento da pesquisa, acarretando vantagens e desvantagens, fazendo-as mais ou menos adequadas de acordo com sua aplicação. As subseções seguintes detalham essas abordagens.

4.3.1. Abordagem virtual

Na abordagem virtual os dados são extraídos diretamente das fontes apenas quando solicitados em uma consulta. Isso traz a vantagem de que os dados estão sempre atualizados, tornado sua aplicação atraente nos casos em que os dados mudam constantemente e quando existe um grande número de bases de dados. Porém, se torna ineficiente quando fontes externas estiverem inacessíveis, além de ter um custo de processamento maior durante a pesquisa, ocasionando tempos de consulta maiores.

Uma arquitetura que implementa a abordagem virtual é a arquitetura baseada em mediadores [27]. Um mediador tem como função prover a visão global do Sistema de Integração de Dados. Os mediadores são utilizados como uma camada intermediária entre a consulta do usuário e as fontes de dados. Geralmente se cria um mediador para cada domínio de interesse e associa-se este mediador às fontes específicas do determinado domínio. Outro conjunto de componentes formador dessa arquitetura é o conjunto de tradutores, também

chamados de adaptadores ou *wrappers*. Cada tradutor tem a função de converter a consulta vinda das camadas superiores na linguagem específica da sua fonte de dados e de converter os dados das fontes para um modelo de dados comum. A Figura 4.10 apresenta uma ilustração típica da abordagem virtual de integração de dados utilizando a arquitetura baseada em mediadores.

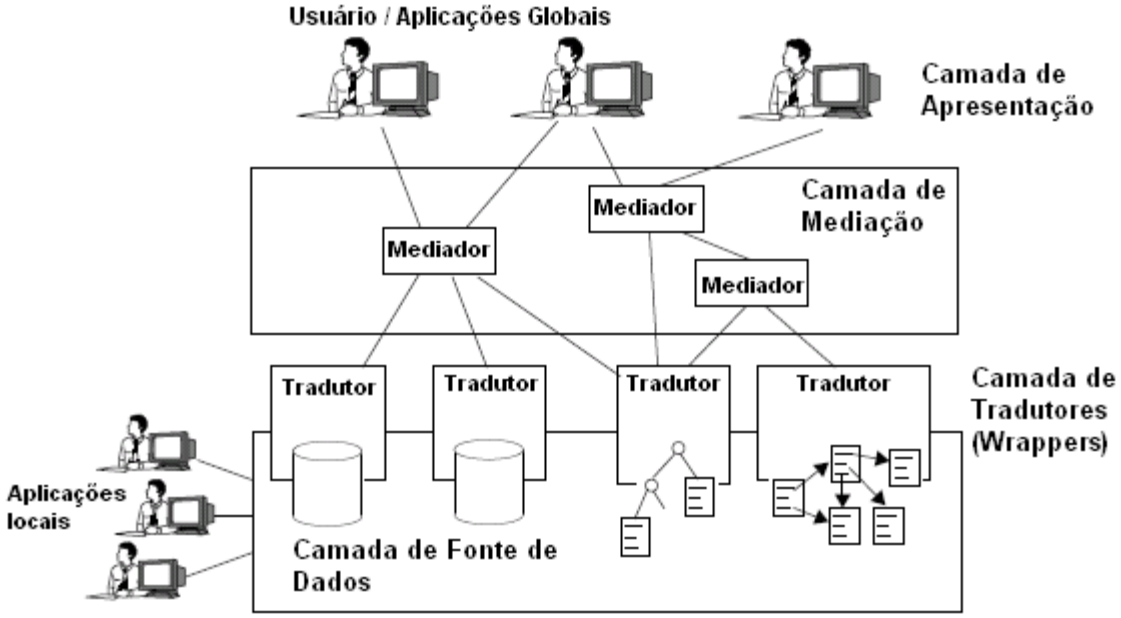


Figura 4.16 – Arquitetura baseada em mediadores. (extraído de [24])

Na camada de apresentação está a interface com o usuário. Uma consulta é solicitada nessa camada e enviada à camada de mediação. Na camada de mediação um conjunto de mediadores forma uma rede que provê o serviço de enviar a consulta às fontes de dados apropriadas. Antes da consulta chegar a cada uma das fontes de dados, ela é traduzida pelo tradutor particular de cada fonte, que então a envia à sua fonte de dados. Consultados os dados nas fontes, estes são retornados e enviados novamente a cada um dos tradutores, que então retornam os dados à camada de mediação. Agora os mediadores centralizam os dados em uma visão unificada. Finalizando, esses dados que compõem a resposta à consulta são enviados à camada de apresentação, onde são visualizados pelo usuário.

4.3.2. Abordagem materializada

Na abordagem materializada os dados relevantes são recuperados, integrados e armazenados em um repositório central, no qual são processadas as consultas feitas ao sistema de integração sem haver acessos diretos às fontes de dados. Nesse caso, há a desvantagem visível de que a base local materializada não está sempre atualizada, sendo necessárias manutenções para que esta base esteja consistente e atualizada com as bases originais. Por

outro lado, a consulta pode ser feita mesmo que todas as fontes de dados, exceto a base materializada, estejam indisponíveis. Mas, a principal vantagem desse tipo de abordagem é em relação ao custo de processamento da consulta comparado à abordagem virtual, tendo a abordagem materializada tempos de consulta menores. Assim, sua aplicação é interessante quando se busca por um bom desempenho no tempo de resposta das consultas e não é preocupante o grau de atualização dos dados das respostas. Também, quando constantes consultas são feitas sobre dados históricos e não atualizáveis.

Um tipo de arquitetura que implementa a abordagem materializada é a arquitetura de DW (*Data Warehouse*) [28]. Uma arquitetura característica de DW é mostrada na Figura 4.11.

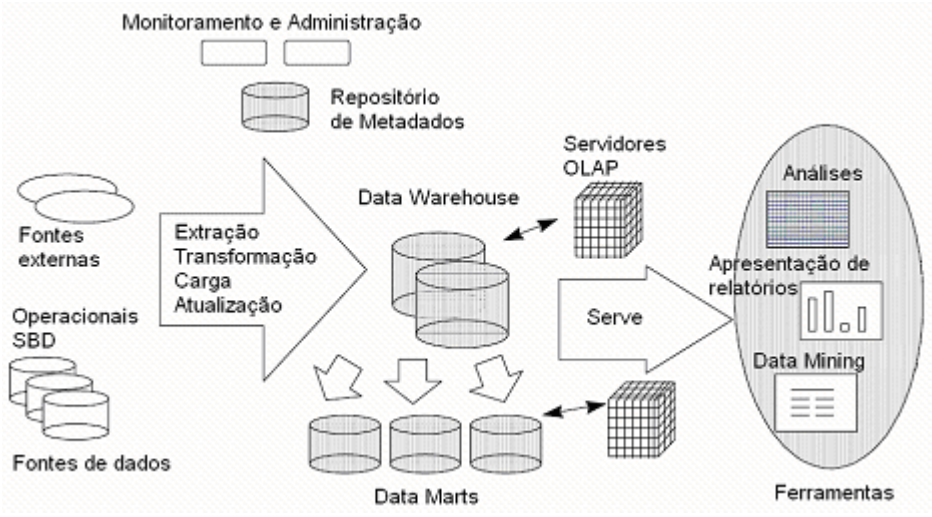


Figura 4.17 – Arquitetura característica de um Data Warehouse. (extraída de [28])

Nessa arquitetura nota-se a presença de *Data Marts,* que são unidades lógicas menores contendo subconjuntos específicos de informações do DW. Também inclui servidores OLAP (*On-Line Analytical Processing*), que são ferramentas capazes de navegar pelos dados, realizando pesquisas e apresentando as informações. Outros componentes são as ferramentas de análises, *Data Mining*, apresentação de relatórios etc, que auxiliam na extração de informações dos dados do DW. Os repositórios de metadados são importantes no gerenciamento do DW ao converter os dados em informações para o negócio, compreendendo informações como: origem dos dados, regras de transformação, formatos de dados, sinônimos etc. Por fim, existem ferramentas para o processo de carga dos dados das fontes operacionais para o DW. Nesse processo, ferramentas são utilizadas para extração, transformação, carga e atualização dos dados.

4.3.3. Abordagem híbrida

Para extrair as vantagens das duas abordagens anteriores, a abordagem híbrida implementa ambas as abordagens em um único sistema, tendo parte dos dados em uma fonte materializada e parte dos dados nas próprias fontes originais. Entretanto, há um aumento considerável na complexidade de todas as partes do projeto, que devem ser avaliados no momento da escolha de qual abordagem (virtual, materializada ou híbrida) será utilizada para integrar as fontes de dados.

Para melhor utilizar as vantagens e minimizar as desvantagens das abordagens virtual e materializada, é necessário que os dados materializados sejam os que raramente sofrem atualizações, como dados históricos. Também os dados das consultas mais freqüentes podem ser materializados para melhor desempenho do sistema.

Para implementar essa abordagem, utilizam-se arquiteturas das abordagens virtual e materializada, no caso a arquitetura de mediadores e de DW respectivamente. A Figura 4.12 mostra uma arquitetura híbrida. Essa é a arquitetura de um *framework*, proposta por Alasoud, Haarslev e Shiri [29]. Tal *framework* aceita qualquer fonte de dados: ontologias, dados da *Web*, ou qualquer fonte estruturada ou semi-estruturada. Além disso, a linguagem OWL DL é usada como formalismo para a visão integrada e fontes de ontologias.

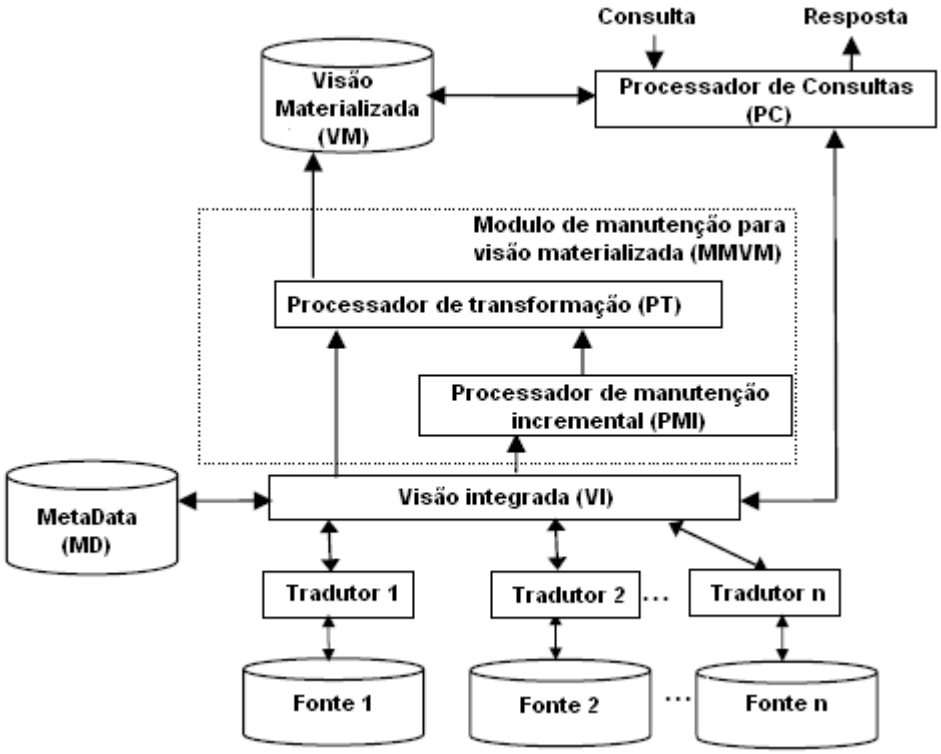


Figura 4.18 – Arquitetura de um framework que utiliza a abordagem híbrida. (extraído de [29])

Esse *framework* é baseado na visão integrada (VI) e em um conjunto de tradutores. O processador de transformação (PT) transforma os dados das fontes para a visão materializada (VM). O processador de manutenção incremental (PMI) determina quais dados do VM serão atualizados. O processador de consultas (PC) determina uma ou ambas as visões para se realizar a consulta. O metadados (MD) é um repositório para o mapeamento dos conceitos, regras e indivíduos, usado pela VI e pelas fontes de dados.

4.4. O uso da ontologia em sistemas de integração de dados

Para promover a integração de dados é necessário um entendimento comum e não ambíguo dos termos e conceitos utilizados nas fontes heterogêneas. Nesse sentido, a ontologia vem sendo utilizada na implementação de Sistemas de Integração de Dados, como base para solução dos problemas causados pela heterogeneidade lógica das fontes. A ontologia, ou um conjunto delas, fornece semântica aos dados provendo o entendimento comum e não ambíguo dos termos e conceitos, servindo como visão homogênea do domínio. Podem existir outros papéis adicionais ao uso da ontologia nos SIDs, como a utilização desta como modelo de consulta [32]. Nesse caso, tem-se a vantagem de uma consulta mais intuitiva ao usuário, mas o limita ao esquema do modelo de consulta.

Existem três arquiteturas de utilização de ontologia nos SIDs [31]:

- Arquitetura de ontologia única: possui uma única ontologia chamada de ontologia global, que provê um vocabulário compartilhado para a especificação da semântica do domínio. Todas as fontes de dados são relacionadas à ontologia global, criando-se um modelo independente para cada fonte e mapeando os objetos do modelo da fonte com os do modelo da ontologia global. Essa arquitetura pode ser aplicada na integração de fontes que compartilham um mesmo escopo;
- Arquitetura de múltiplas ontologias: cada fonte de dados é descrita por sua própria ontologia. Para estabelecer o entendimento comum entre as fontes, sem um vocabulário comum, é feito o mapeamento inter-ontologias. Nesse mapeamento podem ser considerados os diferentes escopos do domínio. Entretanto, a falta do vocabulário comum torna tal mapeamento bastante difícil de ser definido, podendo ocorrer muitos problemas de heterogeneidade lógica;
- Arquitetura híbrida: reúne as características das duas arquiteturas anteriores, procurando superar os inconvenientes de ambas. Nela, cada fonte é descrita por sua

própria ontologia, porém, as ontologias são construídas sob um vocabulário compartilhado. Dessa forma, o mapeamento inter-ontologias fica simples e escopos diferentes podem fazer parte do domínio a ser integrado.

A Figura 4.13 apresenta as três arquiteturas. A Tabela 4.2 sumariza os benefícios e inconvenientes dessas arquiteturas.

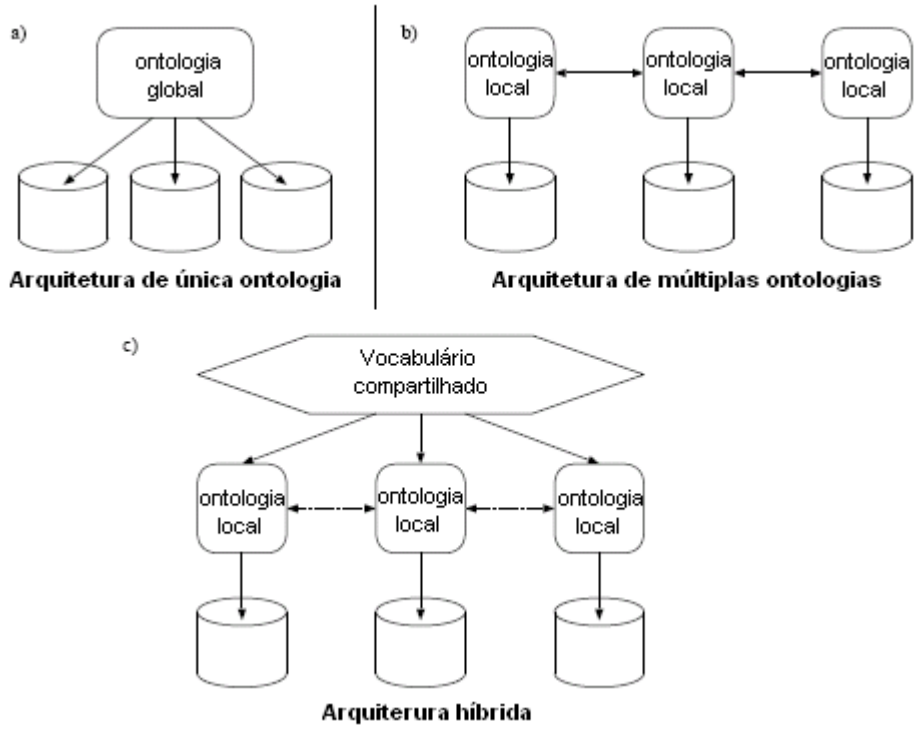


Figura 4.19 – As arquiteturas para utilização de ontologia nos SIDs. (extraído de [31])

	Única ontologia	**Múltiplas ontologias**	**Híbrida**
Esforço de implementação	Direto	Custoso	Razoável
Heterogeneidade semântica	Visão similar de um domínio	Suporta visões heterogêneas	Suporta visões heterogêneas
Adição/remoção de fontes	Precisa de algumas adaptações na ontologia global	Provendo uma nova ontologia fonte; relacionar às outras ontologias	Provendo uma nova ontologia fonte
Comparação de múltiplas ontologias	-	Difícil devido à falta de um vocabulário comum	Simples, porque as ontologias usam um vocabulário comum

Tabela 4.2 – Benefícios e inconvenientes das diferentes arquiteturas para utilização de ontologias nos SIDs. (extraído de [31])

Para estabelecer a conexão da ontologia com outras partes do sistema é realizado um mapeamento. Independente que arquitetura empregada seja a de única ontologia, múltiplas ontologias ou híbrida, é necessário relacionar a(s) ontologia(s) com as fontes de

dados do sistema. Conforme Wache e colegas [31], diferentes abordagens são utilizadas para estabelecer a conexão entre ontologia e fonte de dados, sendo elas:

- Estrutura Semelhante: uma forma direta para conectar a ontologia com o esquema da base de dados é simplesmente produzir uma cópia um-para-um da estrutura da base de dados e codificá-la em uma linguagem que faça o raciocínio automático possível. A integração é executada então na cópia e pode facilmente ser trilhado por trás dos dados originais;
- Definição de Termos: para tornar claro o significado dos termos em uma base de dados, uma cópia da estrutura não é suficiente. Definição de Termos é um mapeamento que usa a ontologia para definir termos de uma base de dados ou esquema de base de dados. Essas definições não correspondem à estrutura da base de dados, elas são somente relações do termo com a informação que o define. O termo pode ser definido por um conjunto de regras. Entretanto, na maioria dos casos, os termos são descritos por definição de conteúdo;
- Enriquecimento de Estrutura: combina as duas abordagens anteriores, sendo a abordagem mais comum. É construído um modelo lógico que representa a estrutura da base de dados e contém definições de conceitos adicionais;
- Meta-Anotações: baseada no uso de meta-anotações para adicionar informações semânticas às fontes de dados. Essa abordagem está se tornando proeminente com a necessidade de integrar informações presentes na WWW (*World Wide Web*), onde anotações é uma forma natural de adicionar semântica.

A Figura 4.14 expõe um exemplo para cada uma das abordagens apresentadas anteriormente de mapeamento entre ontologia e fonte de dados.

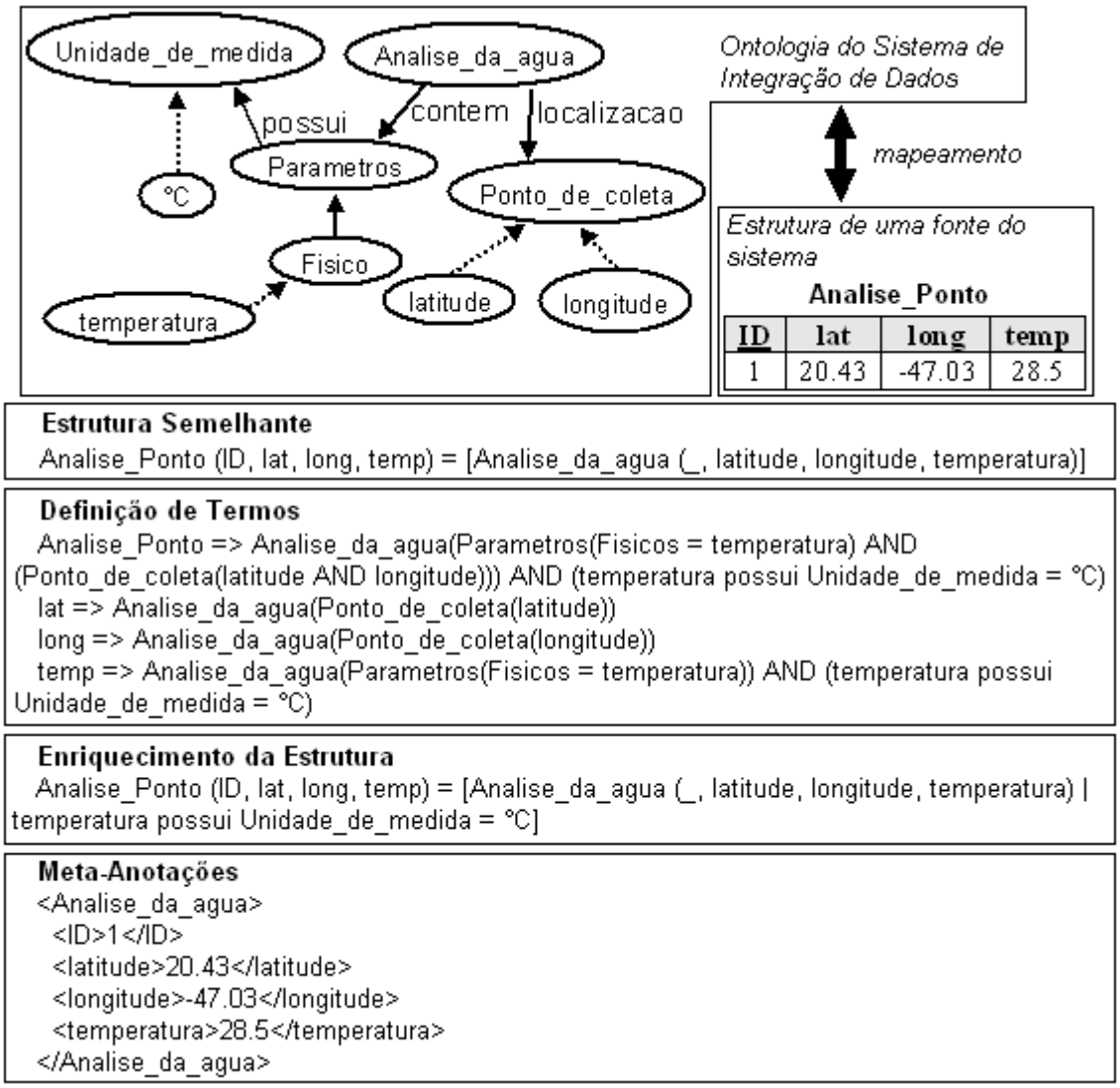


Figura 4.20 – Abordagens de mapeamentos entre ontologia e fonte de dados.

O mapeamento que envolve diferentes ontologias (mapeamento inter-ontologias), usadas em sistemas com arquitetura de múltiplas ontologias ou híbrida, é um problema bem conhecido e discutido na área de Engenharia de Conhecimento e também comentado brevemente por Wache e colegas [31]. Mesmo sendo mapeamentos importantes em sistemas com mais de uma ontologia, não são comentados com detalhes neste trabalho, uma vez que o DISFOQuE utiliza a arquitetura de única ontologia.

Para exemplificar SIDs baseados em ontologia, a seção seguinte comenta alguns, com arquiteturas, mapeamentos e utilização diversificados.

4.5. Exemplos de SIDs baseados em ontologia – alguns dos principais projetos

Há tempos a ontologia vem sendo estudada e empregada na implementação de alguns Sistemas de Integração de Dados. É crescente o número de SIDs que adotam a ontologia para solucionar seus problemas de heterogeneidade semântica. Assim, muitos sistemas são encontrados como referência sobre esse assunto, tendo nesta seção alguns exemplos dos principais projetos.

4.5.1. SIMS

O sistema SIMS (*Services and Information Management for decisions Systems*) utiliza a abordagem de ontologia global, possuindo uma única ontologia como visão do domínio para integrar fontes de dados heterogêneas. Foi desenvolvido para sistemas com fontes de dados dinâmicas, facilitando as alterações destas fontes.

A ontologia global do modelo SIMS é chamada de modelo do domínio e tem o papel de descrever o domínio no qual as informações se encontram. Também descreve o conteúdo e estrutura das fontes de dados, tendo um modelo para cada fonte incorporada ao sistema. Tal modelo contém a linguagem de consulta, a localização da fonte na rede, seu tamanho estimado, a freqüência de atualização, entre outras informações, além de descrever os campos das fontes em termos do modelo de domínio.

A Figura 4.15 apresenta a arquitetura de funcionamento do sistema SIMS, baseado na ontologia global (modelo de domínio).

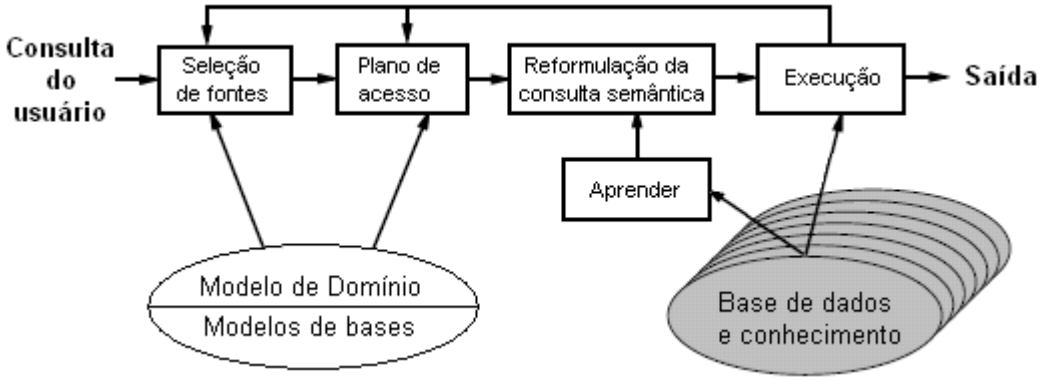


Figura 4.21 – Arquitetura do sistema SIMS. (extraído de [32])

4.5.2. TAMBIS

TAMBIS (*Transparent Access to Multiple Bioinformatics Information Sources*) é um Sistema de Integração de Dados em bioinformática e biologia molecular. Sua ontologia global permite o acesso transparente às fontes de dados desse domínio.

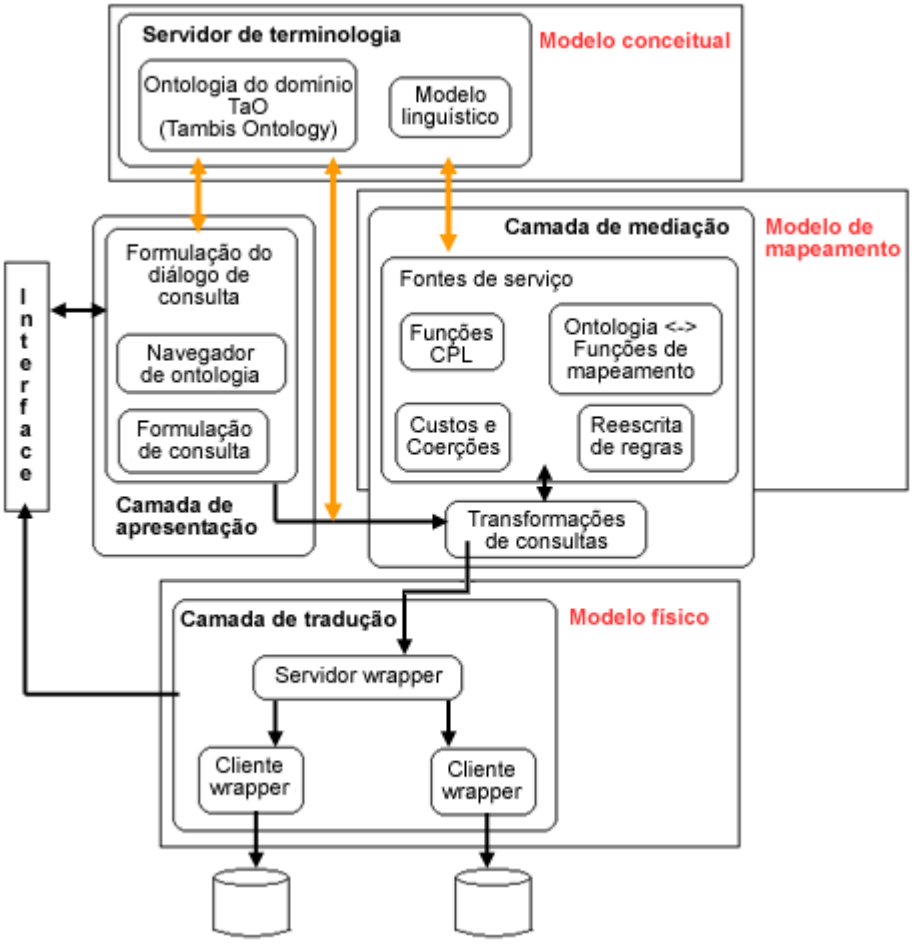


Figura 4.22 – Arquitetura de componentes do sistema TAMBIS. (extraído de [33])

Conforme visto na Figura 4.16, sua arquitetura é composta por cinco componentes principais, organizados em um modelo de mediador e tradutores clássico de três camadas: apresentação, mediação e tradução. O servidor de terminologias armazena e gerencia a ontologia de domínio e o modelo lingüístico, sendo extensivamente utilizado na reformulação e na transformação de consultas. A camada de apresentação é responsável por prover a formulação das consultas. A camada de mediação é responsável por lidar com os mapeamentos, utilizados para a reescrita de consultas. A camada de tradução trata dos tradutores (*wrappers*), provendo acesso às fontes e transformação entre modelos.

4.5.3. ONTOBROKER

O sistema Ontobroker adota a abordagem de múltiplas ontologias. O sistema utiliza ontologias para realizar consultas na Internet. O principal objetivo é extrair, inferir e gerar metadados específicos de domínios para integrar páginas *Web*, utilizando ontologias que refletem consensos entre grupos de usuários *Web*. O Ontobroker utiliza RDF/RDFS para representar seus metadados anotados nas páginas *Web* e utiliza a linguagem F-Logic na máquina de inferência, para responder consultas baseadas em uma ontologia.

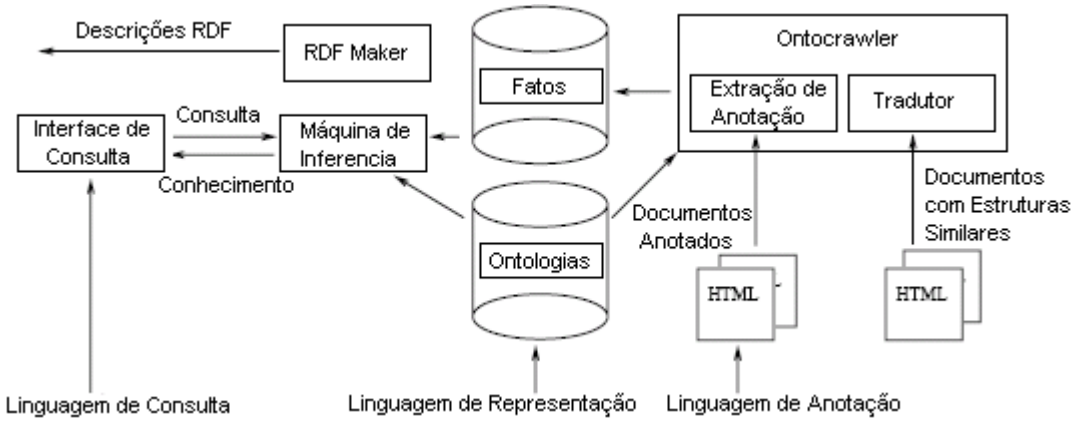


Figura 4.23 – Arquitetura do sistema Ontobroker. (extraído de [34])

Conforme visto na Figura 4.17, a arquitetura do sistema é composta de três elementos principais: uma interface de consultas do usuário, uma máquina de inferência para obter respostas e um agente inteligente utilizado para coletar dados da *Web*. Existem ainda os componentes Ontocrawler, que extraem conhecimento formal das páginas HTML. Também o RDF-Maker, que explora a máquina de inferência e gera uma representação RDF da informação que pode ser inferida na ontologia.

4.5.4. KRAFT

O sistema KRAFT (*Knowledge Reuse And Fusion/Trasformation*) integra fontes heterogêneas usando agentes de softwares para realizar o processo de consultas. Três tipos de agentes são utilizados: os tradutores (*wrappers*) que fazem as ligações com as fontes de dados externas, os facilitadores que proporcionam a comunicação entre agentes e os mediadores, responsáveis pelo processamento interno do conhecimento obtido por outros agentes e facilitadores. Além desses agentes, existem os agentes de usuário que permitem acesso dos usuários às informações por algum tipo de interface de usuário. A Figura 4.18 apresenta essa visão conceitual da arquitetura do KRAFT.

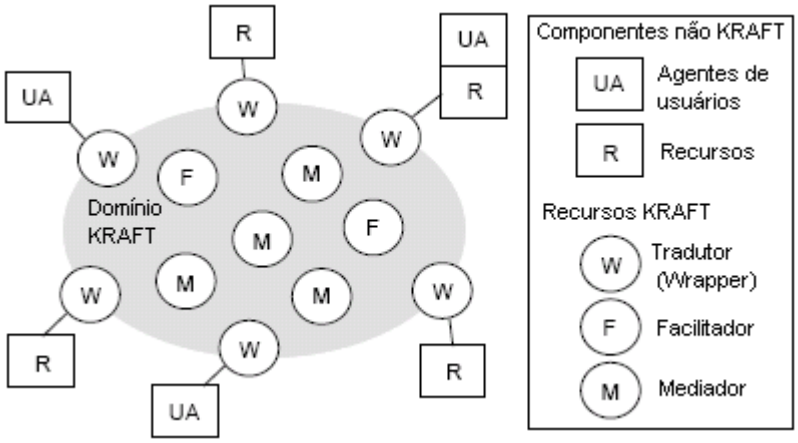


Figura 4.24 – Visão conceitual da arquitetura do KRAFT. (extraído de [35])

Para resolver o problema da heterogeneidade, as mensagens trocadas pelos agentes do sistema são expressas utilizando os termos de uma ontologia compartilhada, que define formalmente a terminologia do domínio. Porém, cada fonte de dados possui sua ontologia local, mapeada para a ontologia compartilhada, o que caracteriza o sistema KRAFT como sendo de arquitetura híbrida.

A Tabela 4.3 sintetiza as características de cada um dos SIDs baseados em ontologia apresentados nesta seção.

SID	Arquitetura em relação à ontologia	Arquitetura de funcionamento	Mapeamento entre ontologia e fonte	Utilização
SIMS	Ontologia global	Mediador e tradutores	Estrutura semelhante	Geralmente integrar fontes estruturadas
TAMBIS	Ontologia global	Mediador e tradutores	Definição de termos	Integrar fontes do domínio de bioinformática
ONTOBROKER	Múltiplas ontologias	Baseada em meta-anotações	Meta-anotações	Integrar páginas *Web*
KRAFT	Híbrida	Mediadores e tradutores (agentes)	Enriquecimento de estrutura	Integrar fontes estruturadas

Tabela 4.3 – Características de alguns dos principais SIDs baseados em ontologia.

4.6. Conclusão do capítulo

Os Sistemas de Integração de Dados proporcionam uma visão única e transparente ao usuário das fontes de dados disponíveis para consultas. Dessa forma, uma série de vantagens são obtidas ao se realizar a tarefa de busca e integração dos dados. Entretanto, existem desafios a serem enfrentados na integração dos dados, sendo a

heterogeneidade semântica e estrutural das fontes um dos principais desafios. Para auxiliar na solução desse problema, a ontologia vem sendo utilizada como base de Sistemas de Integração de Dados, provendo um vocabulário comum aos termos do domínio.

Assim, com a base teórica vista até então neste trabalho, o capítulo seguinte inicia uma descrição do Sistema de Integração de Dados baseado em ontologia desenvolvido como pilar do trabalho de mestrado, o DISFOQuE. Como objetivo inicial esse sistema tem como finalidade integrar dados do domínio de bacias hidrográficas com escopo na análise das mesmas. Entretanto, outros domínios com fontes estruturadas ou semi-estruturadas podem ser integrados pelo sistema.

5. O Sistema de Integração de Dados DISFOQuE

O objetivo primário relacionado a toda pesquisa realizada na elaboração deste trabalho é o desenvolvimento de um Sistema de Integração de Dados (SID) baseado em ontologia. Como motivação à construção desse sistema tem-se o domínio de pesquisas realizadas dentro do Instituto Internacional de Ecologia em São Carlos. Esse domínio envolve análises de bacias hidrográficas, principalmente relacionadas à qualidade das águas. Nesse contexto, um SID foi desenvolvido com características especiais para tal domínio, destacando-se o tratamento dos dados espaciais (com localização geográfica) importantes no domínio. Esse sistema é denominado DISFOQuE (*Data Integration System using Fuzzy Ontology-based Query Expansion*). Entretanto, o propósito de utilização do DISFOQuE se expande para outros domínios, de forma geral, que contenham fontes heterogêneas estruturadas ou semi-estruturadas.

Este capítulo inicia a descrição do DISFOQuE. Inicialmente na seção 5.1 emite uma descrição geral que fala sobre as abordagens utilizadas em sua construção. Após, a seção 5.2 comenta sua constituição, descrevendo suas partes e seu fluxo de funcionamento. Prossegue com a seção 5.3 falando sobre o mapeamento semântico realizado. Entrando em um assunto mais técnico, a seção 5.4 expõe as tecnologias utilizadas na implementação do SID. Finalizando, a seção 5.5 explica as soluções realizadas aos desafios da integração, principalmente relacionados às heterogeneidades lógicas das fontes, e a seção 5.6 faz as conclusões finais sobre o capítulo.

5.1. Descrição geral – abordagens utilizadas

Para iniciar a descrição do DISFOQuE serão apresentadas as características que o classificam nas abordagens discutidas anteriormente no capítulo 4, em relação à integração de dados e ao uso da ontologia.

No primeiro caso, em relação à abordagem utilizada para integrar as fontes de dados, utiliza-se a abordagem virtual implementada com a arquitetura de mediadores (ver subseção 4.3.1 Figura 4.10). Segue o modelo típico dessa arquitetura, com três camadas: de apresentação, de mediação e de tradutores, além das fontes de dados, da ontologia (que se relaciona com todas as camadas do sistema) e de um Banco de Dados Geográfico usado na localização dos dados espaciais.

Em relação ao uso da ontologia, emprega-se a abordagem de única ontologia (ver seção 4.4 Figura 4.13a). Essa ontologia global serve como um modelo do domínio, provedor de um vocabulário único e de regras que relacionam os termos e são utilizadas nas inferências das consultas. Entretanto, seu uso é transparente ao usuário, não sendo consultada por este para formular as consultas.

Outras características do sistema se referem às suas fontes de dados. O sistema integra fontes heterogêneas, de naturezas bem diversas em relação aos dados que armazenam, as estruturas que utilizam, aos modelos de dado e interfaces de acesso. Desse modo, todos os tipos de heterogeneidades comentadas no capítulo 4 (ver subseção 4.2.4) são encontradas e tratadas pelo sistema. Entretanto, as fontes integradas pelo DISFOQuE não podem ser totalmente desestruturadas, como documentos de textos. Assim, devem denotar formalmente e claramente, sem ambigüidade, cada dado contido em sua estrutura. Dessa forma, apenas fontes estruturadas (Bancos de Dados Relacionais ou Orientados a Objetos) ou fontes semi-estruturadas (arquivos XML) são integradas pelo DISFOQuE.

Outra característica das fontes do sistema é que elas são totalmente autônomas, sendo construídas sem qualquer restrição ou modelo que tivesse como objetivo incluí-las em um sistema de integração, ou seja, em nenhum momento foram projetadas para fazer parte de uma integração de fontes. Cada fonte foi desenvolvida isoladamente e trabalha de forma autônoma.

5.2. Descrição da arquitetura do DISFOQuE e seu fluxo de funcionamento

Dadas as características gerais do DISFOQuE, uma descrição mais detalhada da sua constituição será apresentada nesta seção. A Figura 5.1 apresenta a arquitetura desse SID. As setas indicam o fluxo da consulta (1, 2, 3, 4, 5, 6, 6B, 7, 8, 9) e os relacionamentos existentes entre os componentes (R1, R2, R3, R4).

No caso do fluxo, é iniciado com a elaboração da consulta pelo usuário (1), passando pela normalização do vocabulário (2) e expansão da consulta (3), distribuição dela aos tradutores (4) e tradução para cada fonte do sistema (5), retorno dos dados das fontes (6), possível processamento para localização geográfica (6B), envio dos dados normalizados semanticamente para sua integração (7), processamento da integração dos dados e envio dos

dados integrados à interface de visualização (8) e finalmente a visualização dos resultados da consulta pelo usuário (9).

Já os relacionamentos (R1, R2, R3 e R4) envolvem a ontologia, que é a base do sistema de integração, com os demais componentes. O sentido das setas se dá do interferente a quem sofre a interferência. A relação R1 envolve a passagem de informações da ontologia para a camada de formulação de consulta. Na interface dessa camada, a ontologia é examinada para colocar a consulta do usuário em um vocabulário comum e para inferências em relação aos termos da consulta. Também nessa relação, o módulo do sistema FOQuE solicita informações à ontologia para verificar possíveis reescritas da consulta baseadas nas relações semânticas entre os termos. A relação R2 indica a visão global no mediador baseada no vocabulário comum provido pela ontologia. A relação R3 representa a visão que os tradutores têm da ontologia na construção dos mapeamentos semânticos. Já a relação R4 mostra a visão da ontologia sob o BD Geográfico, que lhe fornece dados para criação de novos termos e inferências para o estabelecimento de relações entre estes termos.

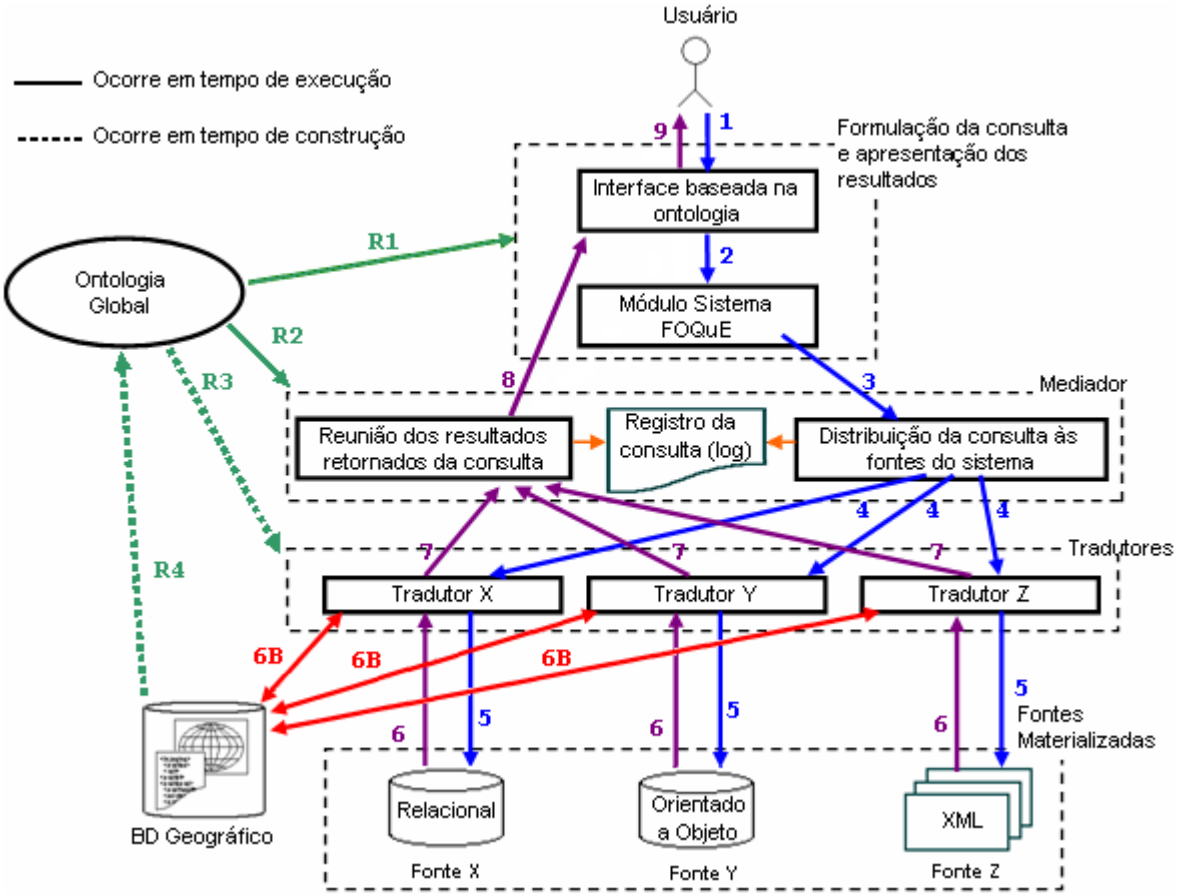


Figura 5.25 – Arquitetura do DISFOQuE.

Uma descrição de cada um dos componentes do DISFOQuE é realizada a seguir:

i. Camada de formulação da consulta e apresentação dos resultados: responsável pela interface entre o usuário e o sistema e pela reescrita da consulta para uma pesquisa mais expressiva semanticamente. Seus dois componentes são:

- Interface baseada na ontologia: onde o usuário formula a consulta e esta é analisada e transcrita em um vocabulário comum. A consulta é realizada em uma caixa de texto, escrevendo os termos do domínio que se deseja consultar (sem vocabulário específico) e suas possíveis restrições (filtros de valores: >, < e =). Faz-se uso dos operadores lógicos E (*AND*) e OU (*OR*) para relacionar os termos da consulta. A ontologia age nessa interface de forma transparente ao usuário, sendo consultada pelo sistema para converter os termos da consulta para um vocabulário comum e fazer inferências na consulta elaborada pelo usuário. Por exemplo, considere a consulta *dbo > 3* escrita por um usuário. Com base no vocabulário da ontologia, o termo *dbo* é convertido em *demanda_bioquimica_de_oxigenio* e em adição, será verificado que o valor *3* necessita de uma unidade de medida específica, solicitando-a ao usuário. Após o usuário inserir a unidade de medida, como *µg/l* que equivale a *0.001* em relação à unidade base, a consulta final é repassada para o sistema como *demanda_bioquimica_de_oxigenio > 0.003*;
- Módulo Sistema FOQuE: o sistema FOQuE (*Fuzzy Ontology Basic Query Expand*) é resultado do trabalho de mestrado desenvolvido por Cristiane Akemi Yaguinuma [11]. Esse sistema tem o objetivo de reescrever a consulta, expandindo-a conforme a semântica do(s) termo(s) da consulta original expressos na ontologia, utilizando as premissas da lógica difusa (*fuzzy*) inseridas nas relações da ontologia. No DISFOQuE, esse módulo expande a consulta pelas regras de similaridade, proximidade todo-parte e transitividade; muito bem definidas em [11]. Por exemplo, se na consulta o usuário estiver procurando pelo termo *total de sólido dissolvido*, o módulo do FOQuE expande essa consulta, por similaridade, também para *resíduos totais dissolvidos*, uma vez que a ontologia relaciona esses dois termos com um grau de similaridade igual a *0.8* (80%), superior ao grau mínimo de similaridade, supor *0.7*, configurado no sistema pelo usuário.

ii. Mediador: é utilizado como camada intermediária entre a camada das aplicações e a camada das fontes de dados, distribuindo a consulta da camada superior de formulação da consulta aos tradutores e posteriormente centralizando os dados fornecidos pelos

tradutores em uma visão unificada. No DISFOQuE é representado por três partes, que são:

- Distribuição da consulta às fontes do sistema: a primeira tarefa do mediador é distribuir a consulta vinda da camada de formulação de consulta a cada um dos tradutores do sistema. Para isso, possui um registro com a identificação o endereço de rede (IP) e porta de conexão de cada um dos tradutores. Para cada tradutor é gerado um sub-processo (*Thread*) que aguarda os resultados da consulta naquele tradutor;
- Reunião dos resultados retornados da consulta: com o retorno dos resultados enviados pelos tradutores, estes dados devem ser reunidos em uma única estrutura, o que de fato consuma a integração. Para integrar os dados, que podem ser de naturezas diferentes, como por exemplo, dados de *nitrogênio* de uma fonte e de *temperatura do ar* de outra fonte, é aplicada a álgebra booleana, sendo o operador E a intersecção e o operador OU a união dos resultados vindos dos diferentes tradutores;
- Registro da consulta (*log*): durante o processamento da consulta, o mediador gera um registro (*log*) de como esta se procedeu. Esse registro contém qual a consulta realizada, a data e hora de início, os tradutores pesquisados, possíveis falhas de comunicação com tradutores, se a consulta obteve resposta e outras possíveis falhas no sistema.

iii. Camada de tradutores: cada fonte de dados possui seu tradutor, o qual tem como função converter as consultas escritas pela interface e enviadas pelo mediador em consultas específicas da sua fonte de dados; e posteriormente, converter os dados da fonte para um modelo de dados comum. Por exemplo, se a consulta *potencial_hidrogenionico* chega a um tradutor de um Banco de Dados Relacional, o qual contenha os dados para tal consulta inseridos no atributo *ph* da tabela *parametros_analise* com os atributos identificadores *latitude*, *longitude* e *data*, a consulta será convertida para

SELECT latitude AS chave_latitude, longitude AS chave_longitude, data AS chave_data, ph AS potencial_hidrogenionico FROM parametros_analise

Para realizar a conversão da consulta enviada pelo mediador de forma única e escrita sob um vocabulário comum, cada tradutor realiza o mapeamento semântico entre sua fonte de dados e os termos da ontologia. Esse processo de mapeamento semântico é detalhado na seção seguinte.

iv. Camada de fontes materializadas: são as fontes de dados heterogêneas integradas pelo DISFOQuE. Como descrito anteriormente, são fontes distribuídas, heterogêneas, autônomas, estruturadas ou semi-estruturadas. O sistema é flexível em relação às fontes, pois essas podem sofrer alterações em sua estrutura e principalmente permite a inserção de uma nova fonte a ser integrada pelo sistema.

v. Ontologia Global: constitui o modelo do domínio, provendo o vocabulário comum à integração dos dados. A ontologia é escrita na linguagem OWL DL com metadados para utilização das propriedades da lógica difusa providas pelo módulo do sistema FOQuE.

vi. Banco de Dados Geográfico: criado para buscar e integrar os dados com características geográficas, ou seja, com localização espacial, utilizado especialmente no domínio de bacias hidrográficas, principal domínio de interesse deste trabalho. Armazena dados espaciais [36] com os limites das bacias, os cursos de água (rios, córregos, ribeirões etc), os limites estaduais e municipais, as relações entre as bacias (quanto uma bacia está contida dentro de outra), entre os rios e as bacias (quanto um rio está contido dentro de uma bacia) e entre bacias e municípios (quanto uma bacia está contida dentro de um município). Como exemplo de utilização, suponha a consulta do usuário *potencial_hidrogenionico AND rio = Tiete*, ou seja, os dados de *potencial hidrogenionico* do *rio Tietê*. Considere o caso em que uma fonte de dados não possui qualquer semântica para identificar *rio*, mas possui valores para as coordenadas geográficas (latitude e longitude) dos pontos de coleta de *potencial_hidrogenionico*, então os dados retornados da consulta a este parâmetro têm suas coordenadas submetidas a um processamento no BD Geográfico, o qual retorna os pares de coordenadas localizados dentro do *rio Tietê*. Também um relacionamento envolve esse BD Geográfico com a ontologia. Nesse caso, é possível, analisando-se o BD Geográfico, criar na ontologia global instâncias de rios e bacias hidrográficas e relacioná-las, indicando quais as sub-bacias de uma bacia, quais os rios contidos em uma bacia, quais as bacias contidas em um município etc.

No capítulo 6 será exemplificado o processamento de uma consulta do domínio de bacias hidrográficas no DISFOQuE. Dessa forma, a seqüência do fluxo da consulta será detalhada e melhor visualizada em uma continuidade das etapas realizadas durante tal fluxo.

5.3. O mapeamento semântico

Dois fatos se confrontam nos tradutores do sistema. Primeiro: a consulta que chega aos tradutores do sistema possui termos contidos em um vocabulário comum, obtido pela ontologia. Segundo: cada fonte de dados, relacionada a seu próprio tradutor, nomeia e estrutura seus elementos de maneira própria. E para tratar esses dois fatos, adaptando a consulta única do sistema de integração a uma consulta respectiva na linguagem de sua fonte de dados específica, cada tradutor deve relacionar os termos da sua fonte aos termos semanticamente iguais na ontologia. Assim, essa ação é chamada de mapeamento semântico.

Para discutir o assunto de como o DISFOQuE faz o mapeamento semântico, será usada como exemplo uma fonte no modelo relacional, com linguagem SQL, nomeada de *EX*, representada na Figura 5.2 pelo seu modelo lógico.

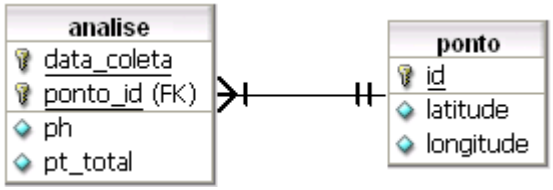


Figura 5.26 – Modelo lógico da fonte de dados *EX*.

Iniciando a discussão, uma consulta simples a *potencial_hidrogenionico* é realizada, estando nos termos do vocabulário comum da ontologia, chegando ao tradutor da fonte *EX*. Um mapeamento simples e direto é relacionar o termo *potencial_hidrogenionico* ao atributo *ph* da entidade *analise*, traduzindo a consulta para:

SELECT analise.ph AS potencial_hidrogenionico FROM analise

Entretanto, a consulta sendo realizada dessa forma simples, apenas com o retorno dos dados do atributo *ph*, levanta a questão de: como esses dados serão utilizados pelo sistema e integrados aos dados retornados das demais fontes? Para responder essa questão uma outra é atribuída: como identificar cada instância dos dados retornados e assim, esses serem relacionados uns com os outros, verificando se são integrantes da mesma instância?

Possivelmente, a primeira forma que se imagina para identificar cada instância é com os dados dos atributos identificadores da própria entidade em que pertence a instância. Dessa forma, continuando a consulta de exemplo *potencial_hidrogenionico*, esta seria traduzida para:

SELECT analise.ph AS potencial_hidrogenionico, analise.data_coleta AS chave_data, analise.ponto_id FROM analise

Observa-se que o atributo *data_coleta* é mapeado para o termo *data*, semanticamente igual na ontologia, tendo a palavra *chave_* anteposta indicando que tal atributo é um identificador da entidade. Porém, o atributo *ponto_id*, que também é um identificador da entidade, não foi mapeado para nenhum outro termo da ontologia. Isso porque esse não é um atributo relacionado a nenhum objeto do mundo real, sendo chamado de chave substituta (*surrogate key*) [40]. Assim, tal fato torna a identificação válida apenas para a fonte *EX*, mas incompatível para o relacionamento dessas instâncias com as instâncias das demais fontes de dados. Veja por exemplo o caso de uma instância que tem como identificador *id* o valor *1*. Este valor não pode ser relacionado a nenhum outro valor de outra fonte qualquer, mesmo que esta também tenha o atributo identificador chamado *id* com uma instância com valor *1*, pois isso não significa que tais instâncias sejam as mesmas, uma vez que ambos os atributos *id* não se relacionam com objetos semanticamente iguais no mundo real.

Dessa forma, origina-se nessa discussão a questão de utilização de identificadores que representem objetos do mundo real, dentro do domínio modelado pela ontologia. Isso leva à troca das chaves substitutas por chaves naturais (*natural key*) [40]. Nesse caso, seleciona-se a chave candidata, ou seja, um outro atributo ou conjunto de atributos que também identifica a entidade de forma inequívoca, existente como objeto do mundo real. No exemplo da fonte *EX*, a chave natural que substitui o identificador *id* (chave substituta) é o conjunto de atributos *latitude* e *longitude*. Este par de atributos identifica cada instância da entidade *ponto* de forma unívoca. Desse modo, a consulta *potencial_hidrogenionico* enviada ao tradutor da fonte *EX* ficará, considerando o mapeamento entre os elementos que representam e identificam este termo, da seguinte forma:

> *SELECT analise.ph AS potencial_hidrogenionico, analise.data_coleta AS chave_data, ponto.latitude AS chave_latitude, ponto.longitude AS chave_longitude FROM analise, ponto WHERE analise.ponto_id = ponto.id*

Assim, o mapeamento feito pelo tradutor da fonte *EX* ao termo *potencial_hidrogenionico* da ontologia, considerando-se a semântica dos elementos e de seus respectivos identificadores, é realizado em um arquivo XML, conforme mostra o trecho no Quadro 5.1 seguinte:

```
<MAPEAMENTO>
        <TERMO>potencial_hidrogenionico</TERMO>
        <SELECT>analise.ph AS potencial_hidrogenionico</SELECT>
        <FROM>analise, ponto</FROM>
        <WHERE>analise.ph IS NOT NULL AND analise.ponto_id = ponto.id</WHERE>
        <WHERE_PARAMETRO>analise.ph $sinal $valor</WHERE_PARAMETRO>
        <CHAVE>ponto.latitude AS chave_latitude</CHAVE>
        <CHAVE>ponto.longitude AS chave_longitude</CHAVE>
        <CHAVE>analise.data_coleta AS chave_data</CHAVE>
</MAPEAMENTO>
```

Quadro 5.1 – Mapeamento do termo *potencial_hidrogenionico* na fonte EX.

Além dos mapeamentos entre todos os elementos que a fonte de dados possui semanticamente iguais a termos na ontologia, o arquivo XML de mapeamentos contém os dados de conexão com a fonte daquele tradutor e com o BD Geográfico.

Outra questão sobre o mapeamento é em relação às regras especiais, motivadas pelo domínio de bacias hidrográficas, mas que podem servir a outros domínios. Para dar início a essa questão, uma consulta mais elaborada deve ser realizada, como exemplo: *pontencial_hidrogenionico AND rio = Tiete*. Chegando essa consulta ao tradutor da fonte *EX*, este não tem como mapear o termo *rio*, uma vez que sua fonte *EX* não possui um elemento semanticamente igual a *rio*. Entretanto, é possível relacionar as coordenadas, *latitude* e *longitude* da entidade *ponto*, à localização em rios cadastrados no BD Geográfico. Para tal, no mapeamento de um termo, dentro do arquivo XML do tradutor, deve existir um marcador (*tag*) *TIPO_PROCESSAMENTO* que represente qual o tipo de processamento especial o específico termo terá. Para o caso do exemplo acima, o termo *rio* deve passar por um processamento geográfico, feito por funções do BD Geográfico. O Quadro 5.2 mostra o mapeamento realizado para o exemplo do termo *rio*.

```
<MAPEAMENTO>
        <TERMO>rio</TERMO>
        <SELECT></SELECT>
        <FROM>ponto</FROM>
        <TIPO_PROCESSAMENTO>GEOGRAFICO</TIPO_PROCESSAMENTO>
        <CHAVE>ponto.latitude AS chave_latitude</CHAVE>
        <CHAVE>ponto.longitude AS chave_longitude</CHAVE>
</MAPEAMENTO>
```

Quadro 5.2 – Mapeamento do termo *rio* na fonte EX, necessário processamento geográfico.

Outros tipos de processamentos especiais podem ser realizados para o mapeamento semântico entre os termos da ontologia e os elementos da fonte de dados. Como no caso das unidades de medidas, bastante comuns no domínio de bacias hidrográficas. Veja o exemplo da consulta: *fosforo_total > 0.03*. Esta consulta já está normalizada, ou seja, foi reescrita baseando-se na ontologia, com seu vocabulário comum e relações entre os termos.

Assim, o valor *0.03* não contém unidade, pois foi reescrito transformando o valor na unidade em que foi escrito originalmente para um valor proporcionalmente igual na unidade base. Suponha que originalmente a consulta tenha sido escrita colocando-se: *fósforo total > 30 mg/m³*. Na ontologia há uma relação que diz que o valor em *mg/m³* equivale a *0.001* do valor na unidade base. Assim, o valor *30 mg/m³* é transformado para *0.03*. Dessa forma, a consulta será comum para todo o sistema, chegando igual a todos os tradutores. Porém, agora nos tradutores ela será interpretada e transformada observando-se a semântica dos dados armazenados em cada fonte. No exemplo da fonte *EX*, o termo *fosforo_total* é semanticamente igual ao atributo *pt_total* na entidade *analise* e suponha que os valores deste atributo estejam armazenados em *µg/m³*, equivalentes a *0.000001* de sua unidade base. Desse modo, o mapeamento para *fosforo_total* no arquivo XML do tradutor da fonte *EX* fica como mostrado no Quadro 5.3 seguinte.

```
<MAPEAMENTO>
        <TERMO>fosforo_total</TERMO>
        <SELECT>analise.pt_total AS fosforo_total</SELECT>
        <FROM>analise, ponto</FROM>
        <WHERE>analise.pt_total IS NOT NULL AND analise.ponto_id = ponto.id</WHERE>
        <WHERE_PARAMETRO>analise.pt_total $sinal $valor</WHERE_PARAMETRO>
        <TIPO_PROCESSAMENTO>UNIDADE</TIPO_PROCESSAMENTO>
        <FATOR_TRANSFORMA_UNIDADE>0.000001</FATOR_TRANSFORMA_UNIDADE>
        <CHAVE>ponto.latitude AS chave_latitude</CHAVE>
        <CHAVE>ponto.longitude AS chave_longitude</CHAVE>
        <CHAVE>analise.data_coleta AS chave_data</CHAVE>
</MAPEAMENTO>
```

Quadro 5.3 – Mapeamento do termo *fosforo_total* na fonte EX, necessário processamento de unidade de medida.

5.4. As tecnologias empregadas na implementação do sistema

Esta seção trata de uma parte mais técnica do trabalho, expondo as linguagens, sistemas e ferramentas empregadas na implementação do DISFOQuE.

Para a implementação da interface, do mediador e dos tradutores, utilizou-se a linguagem de programação Java [41]. A escolha dessa linguagem para o desenvolvimento do software se justifica por argumentos como:

- É uma linguagem multiplataforma, permitindo que seus códigos sejam interpretados em qualquer sistema operacional e hardware que contenha uma Máquina Virtual Java (*Java Virtual Machine*);

- Facilita a integração com o sistema FOQuE, uma vez que este sistema também é escrito em Java;
- Possibilita a utilização do *framework* Jena [42] para inferências na ontologia escrita em OWL;
- Permite programação concorrente com o emprego da classe Thread [43]. Útil no sistema quando a consulta é distribuída pelo mediador, que cria linhas concorrentes para o envio e recebimento da consulta a cada um dos tradutores.

Em relação ao BD Geográfico do sistema, este é desenvolvido no SGBD objeto-relacional PostgreSQL (Postgres). É um SGBD gratuito, maduro, desenvolvido de forma aberta, robusto, utilizado para armazenar grandes bases de dados com confiabilidade, com distribuições para vários sistemas operacionais, entre outras características vantajosas. Mas, o principal motivo da utilização do PostgreSQL no desenvolvimento do BD Geográfico do DISFOQuE foi a existência da extensão PostGIS, que permite o armazenamento de objetos espaciais e funções para análise e processamento destes objetos. Assim, é possível, por exemplo, armazenar os limites das bacias hidrográficas, formados por polígonos e saber, através de um gatilho (*trigger*) ativo no momento da inserção de um polígono, quais são os outros polígonos de bacias hidrográficas que estão contidos dentro deste, ou seja, quais são suas sub-bacias. Com outra função (*function*) é possível, por exemplo, submeter um par de coordenadas (latitude e longitude de um ponto) a um processamento que retorna em quais bacias hidrográficas o ponto está contido.

Outra linguagem utilizada no desenvolvimento do DISFOQuE foi a XML [44]. Essa linguagem é empregada nos arquivos de mapeamento dos tradutores, no arquivo de registro das consultas e no arquivo de cadastro de tradutores no mediador. A escrita dos arquivos em XML traz uma flexibilidade nas alterações, clareza na escrita e estruturação hierárquica necessária para navegação nos mesmos.

Como dito anteriormente, a ontologia é escrita em OWL DL. Como proposta de padrão da W3C para a *Web* Semântica, ela é freqüentemente utilizada na implementação de ontologias. Esse fato acarreta vantagens como: possibilidade de reutilizar ontologias, existência de ferramentas para criação de ontologias, como a ferramenta Protégé, utilizada para criar as ontologias usadas nos casos de estudos, e utilização de outros códigos como o *framework* Jena, utilizado para fazer inferências nas ontologias criadas.

A Tabela 5.1 resume as tecnologias utilizadas no desenvolvimento do DISFOQuE.

Tecnologia	Empregada em	Motivação
Java	Implementação da interface, mediador, tradutores	Multiplataforma, aproveitamento de códigos, utilização de *frameworks* e programação concorrente
PostgreSQL/ PostGIS	Gerenciamento do BD Geográfico	Robustez. Armazenamento, análise e processamento de objetos espaciais
XML	Criação dos arquivos de mapeamento, registro e cadastro	Representação clara em estrutura hierárquica
OWL	Implementação da ontologia	Ferramentas para criação e inferências. Grande utilização, tornando um padrão

Tabela 5.4 – Tecnologias empregadas no DISFOQuE.

5.5. Soluções realizadas aos desafios de integração

No capítulo 4 (ver seção 4.2) foram vistos alguns desafios enfrentados na tarefa de integração de dados. Agora, esta seção mostra como esses desafios foram tratados DISFOQuE, destacando-se principalmente a questão da heterogeneidade lógica das fontes de dados. Com exceção da distribuição das fontes, todos os outros desafios têm como base de seus cuidados a ontologia.

A distribuição das fontes é solucionada com a utilização das redes de computadores. Atualmente grande parte dos computadores está conectada a alguma rede, principalmente os computadores corporativos, que armazenam as fontes de dados. Outro fato é em relação a essas redes se conectarem a rede Internet, o que vem se tornando um preceito natural, com crescimentos exponenciais através dos anos do número de computadores conectados a Internet. Assim, o padrão de rede que vem se formando utiliza o protocolo TCP/IP, sendo este o protocolo da Internet. Entretanto, o que pode ser considerado ainda um obstáculo maior a esse desafio, uma vez que a conexão dos computadores à rede é quase sempre estabelecida, é a questão da velocidade de comunicação entre os computadores. Apesar do aumento das taxas de transmissão utilizadas nas conexões, é possível que algumas conexões tragam um atraso considerável na consulta aos dados.

Feitas essas observações a respeito das redes de computadores, não obstante o sistema utiliza o protocolo TCP/IP em suas conexões. Uma observação feita é em relação às duas formas de conectar o mediador do sistema as fontes de dados. Na primeira forma, o tradutor da fonte de dados fica na mesma máquina onde se encontra o mediador, mesmo que a fonte esteja em outra máquina. Na segunda forma, o tradutor fica na mesma máquina onde está a sua fonte de dados, podendo dessa forma ficar em uma máquina diferente de onde se

encontra o mediador. Entretanto, em ambas as formas, o modo de conexão entre o mediador e o tradutor se dá do mesmo modo, fornecendo-se o endereço IP e o número da porta do tradutor, armazenados no arquivo de conexão utilizado pelo mediador. A Figura 5.3 apresenta um exemplo da arquitetura física formada pelas conexões entre as máquinas fisicamente distribuídas.

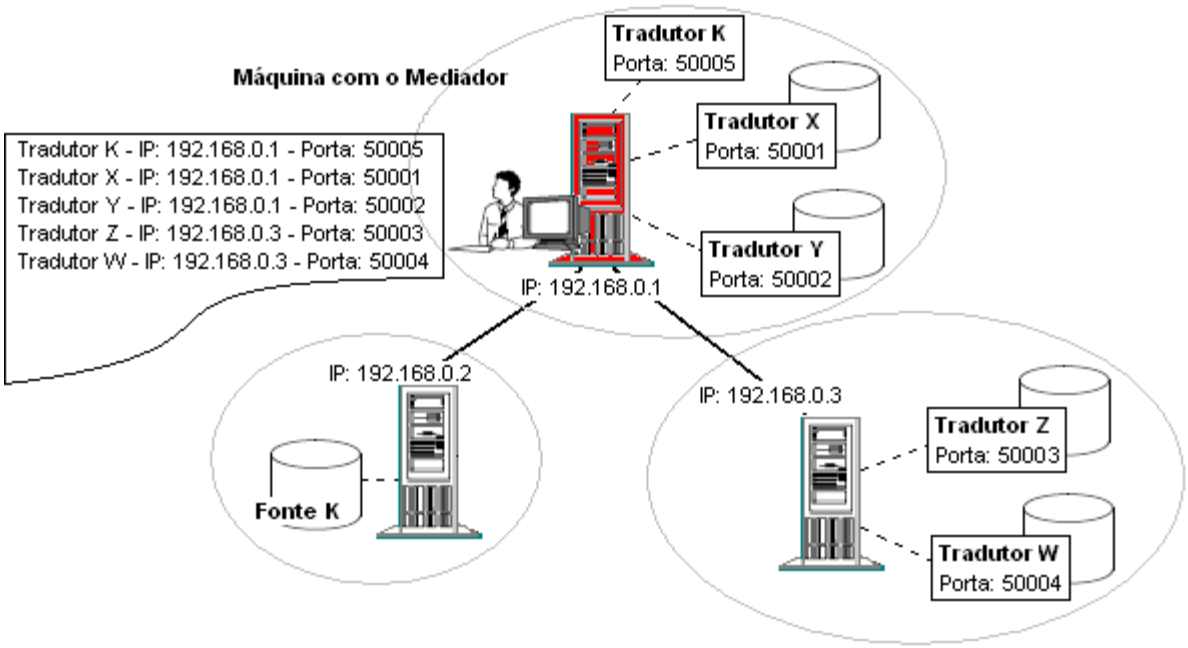


Figura 5.27 – Exemplos da arquitetura física das conexões entre as máquinas do sistema.

Os demais desafios da integração aqui abordados são solucionados basicamente pelo uso da ontologia e mapeamentos feitos nos tradutores. Como no caso da autonomia das fontes, independentes umas das outras e trabalhando isoladamente. Para que uma fonte participe do DISFOQuE é necessário um tradutor para ela. Esse tradutor conterá a conexão com a fonte, ou seja, seu endereço, usuário e senha, caso necessários, e qualquer outra informação pertinente à comunicação com a fonte. Assim, é possível um grau de segurança no qual é permitido criar usuários específicos para acessar a base de dados para uso do sistema. Também é permitido escolher quais informações da fonte serão compartilhadas com o sistema, uma vez que se podem ocultar informações semanticamente iguais a termos na ontologia, mas que não se quer compartilhar, simplesmente não fazendo o mapeamento destas.

No caso da flexibilidade do sistema, também se resume no mapeamento das fontes com os termos da ontologia. Para o acréscimo de uma nova fonte é necessário gerar um tradutor para a fonte. Essa tarefa é facilitada com o uso de componentes gerados para o sistema DISFOQuE como padrões de tradutores específicos para cada modelo de fonte

(relacional, orientado a objetos e XML). Basicamente, é feito o mapeamento dos elementos da fonte com os termos da ontologia no arquivo XML do tradutor e acrescentado o tradutor no arquivo de conexão do mediador. O trabalho maior nesse caso está em se fazer o mapeamento, uma vez que se pode ter uma grande quantidade de elementos na fonte semanticamente iguais a termos na ontologia. Uma questão mais difícil referente à flexibilidade é em relação a mudanças na ontologia. A maioria das alterações na ontologia deve gerar atualizações em cada um dos tradutores, ao menos verificações checando a necessidade de alterações. Todavia, o sistema é bastante flexível, uma vez que se espera o uso de uma ontologia consolidada, construída segundo métodos como o METHONTOLOGY.

Por fim, o desafio geralmente mais custoso, porém a principal motivação do uso da ontologia em Sistemas de Integração de Dados, é a heterogeneidade lógica das fontes. A combinação do vocabulário comum, proveniente da ontologia, e o mapeamento entre os elementos das fontes e os termos deste vocabulário, feito pelos tradutores, é a solução para o problema da heterogeneidade lógica das fontes do DISFOQuE. A seguir é feita uma descrição das soluções para as formas mais comuns de heterogeneidade lógica encontradas no caso de estudo de bacias hidrográficas. Essas descrições serão baseadas em exemplos, considerando-se fontes no modelo relacional. Os mapeamentos, feitos nos arquivos dos tradutores, não serão escritos por completo, apenas os campos relevantes para resolver o conflito em questão. Assim, os conflitos descritos são:

Conflitos de nomes

Considere a Figura 5.4. Observando a figura nota-se a diferença entre os nomes dos vários elementos de cada uma das bases de dados. O mapeamento trata esse conflito respeitando a semântica de cada um dos elementos, sendo estas semânticas iguais, independente do nome.

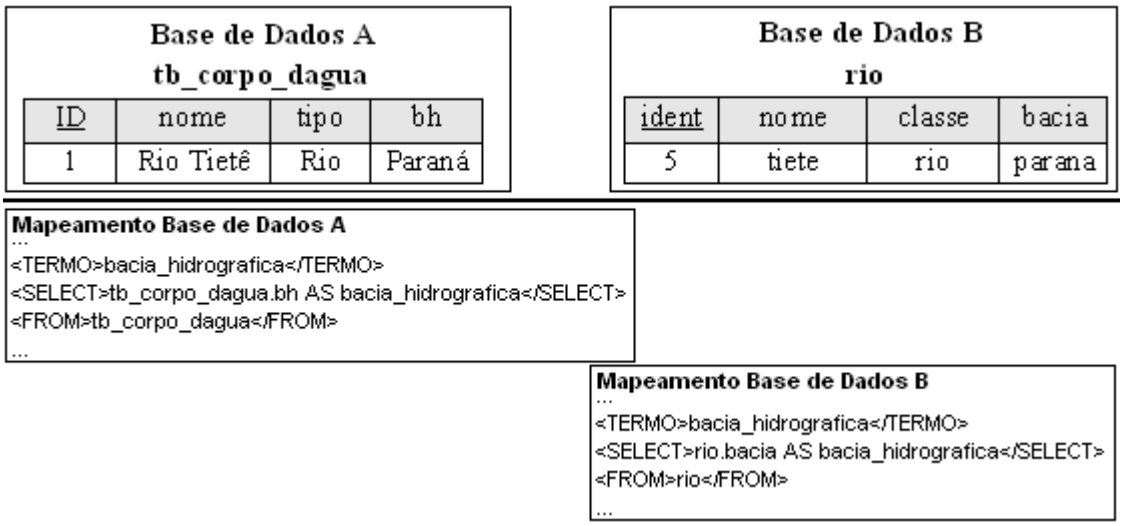


Figura 5.28 – Mapeamentos para conflitos de nomes.

Conflitos de escala

Considere a Figura 5.5. Esta figura mostra um conflito de escala em relação ao atributo *comprimento* das entidades da base de dados *A* e *B*. Na base *A* tal atributo é armazenado em metros, tendo uma relação de 1/1 entre a unidade base. Já na base *B* esse atributo é armazenado em quilômetros, tendo uma relação de 1000/1 entre a unidade base e a unidade dada. Nesse caso, o mapeamento deve conter um tipo de processamento especial para unidades, com fator de 1000 (1000/1) para transformar os valores do atributo comprimento da base *B*.

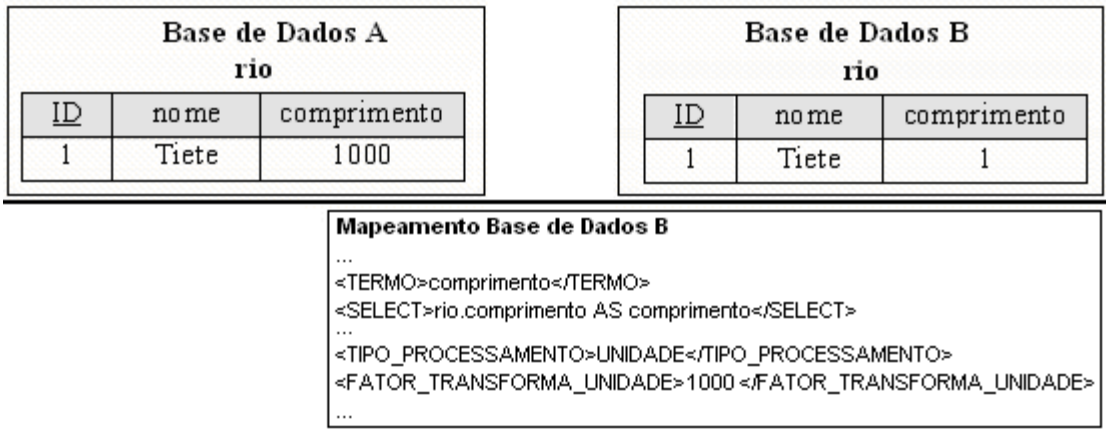


Figura 5.29 – Mapeamentos para conflitos de escala.

Conflitos de identificadores

Considere a Figura 5.6. Há um caso de conflito de identificadores entre a base *A* e *B*. Na primeira existe uma chave substituta *ID*. Já na segunda a chave é natural, composta pelos atributos *latitude* e *longitude*. Entretanto, como regra, em ambos os casos a chave que

irá representar as instâncias da entidade será uma chave natural, sendo que mesmo para a base *A* os atributos *latitude* e *longitude* são os que compõem a chave nos mapeamentos.

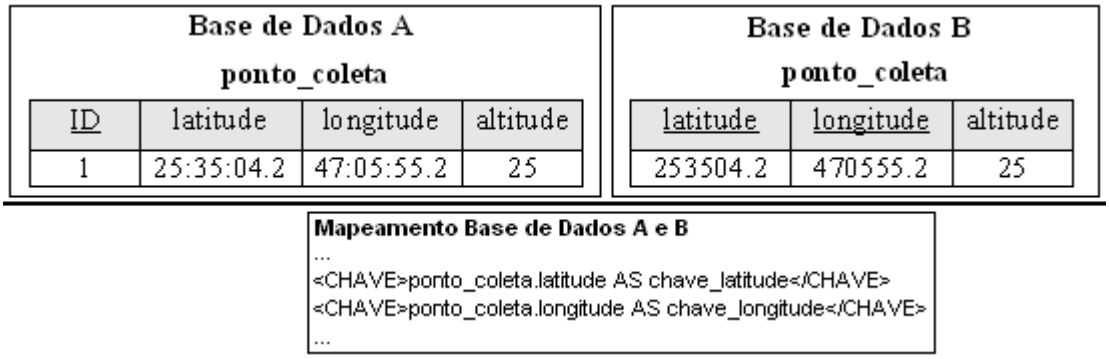


Figura 5.30 – Mapeamentos para conflitos de identificadores.

Conflitos de ausência de dados

Considere a Figura 5.7. O termo *rio* na base de dados *B* é implícito na semântica da entidade *analise_rio_tiete*, sendo que todos os dados ali contidos são referentes ao rio Tietê. Assim sendo, qualquer consulta a *rio*, feita nessa base *B* deverá retornar *Tietê* como resposta. E isso pode ser realizado no mapeamento, mostrado na mesma figura.

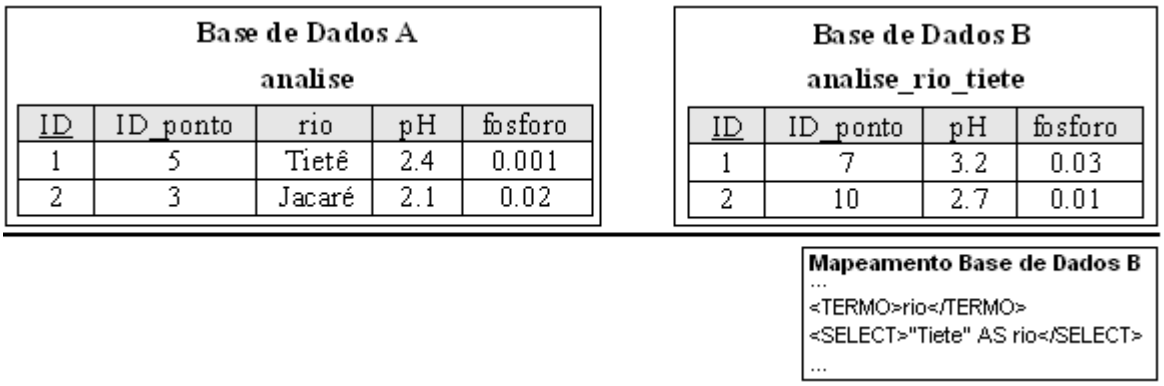


Figura 5.31 – Mapeamentos para conflitos de ausência de dados.

Conflitos de valor e atributo

Considere a Figura 5.8. Essa figura apresenta conflitos de valor e atributo entre as entidades das bases de dados *A* e *B*. Na base *A* o atributo *tipo_uso* armazena os valores *plantio* e *pastagem*, com seus respectivos valores de porcentagem no atributo *porcentagem*. Enquanto que na base *B*, *plantio* e *pastagem* são atributos que armazenam valores das porcentagens em relação a ambos os tipos de uso de solo. No mapeamento desses termos se resolve tal caso de heterogeneidade, como mostra na figura para o termo *plantacao*. Nota-se que no caso da semântica desse termo estar no valor de um atributo, a cláusula WHERE é utilizada para tratar esse conflito, como mostra o mapeamento da base *A*.

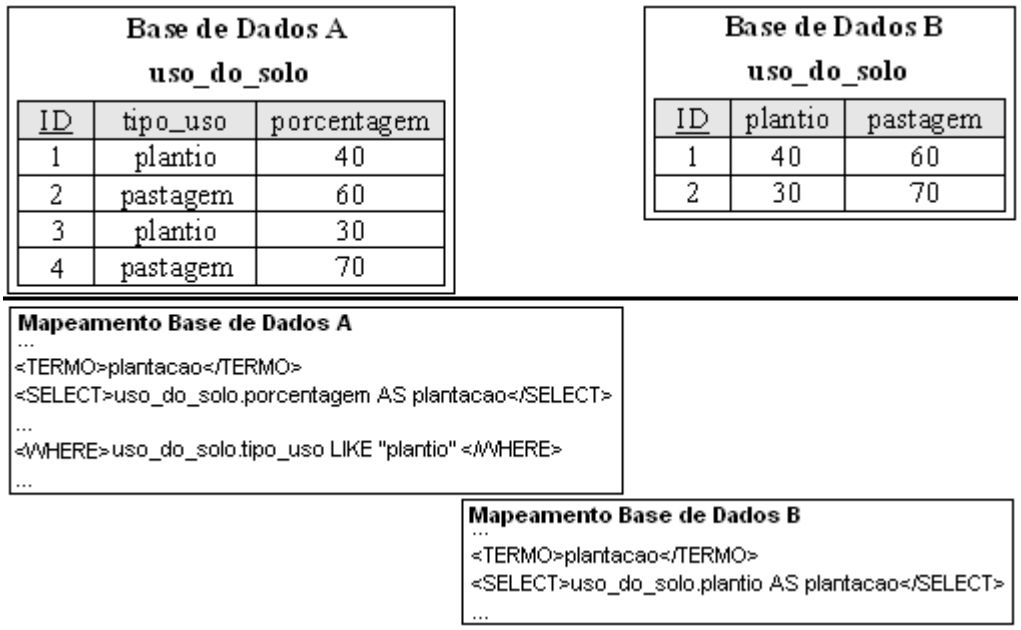


Figura 5.32 – Mapeamentos para conflitos de valor e atributo.

Outros conflitos causados pelas heterogeneidades semânticas e estruturais das fontes também são solucionados utilizando-se essa mesma lógica de respeitar a semântica dos termos que vêem do vocabulário compartilhado da ontologia nos mapeamentos realizados. Alguns desses conflitos até mesmo passam despercebidos, como no caso dos conflitos de generalização, uma vez que os termos são tratados isoladamente no mapeamento. Também os conflitos de representação dos dados são exemplos desses casos despercebidos, uma vez que todos os dados são representados como caracteres (*strings*) no sistema de integração.

Entretanto, dois tipos de conflitos devem ser discutidos em relação à superação deles pelo DISFOQuE. Um desses conflitos é o de precisão. Observa-se na Figura 5.9 um exemplo onde ocorre tal conflito. Os dados até podem ser integrados, mesmo tendo seus valores representados de formas diferentes. Ainda se pode realizar uma transformação dos valores numéricos da base *A* para as faixas de valores armazenadas em *B*, bastando seguir a tabela de relação entre os dados. Todavia, o inverso não é permitido, uma vez que não se sabe o valor numérico exato do dado em meio a uma faixa de valores. Logo, para o sistema retornar uma consulta em que todos os dados estejam em um único formato, a única maneira é converter os dados numéricos para dados de uma faixa de valores, mas esta pode não ser a solução desejada.

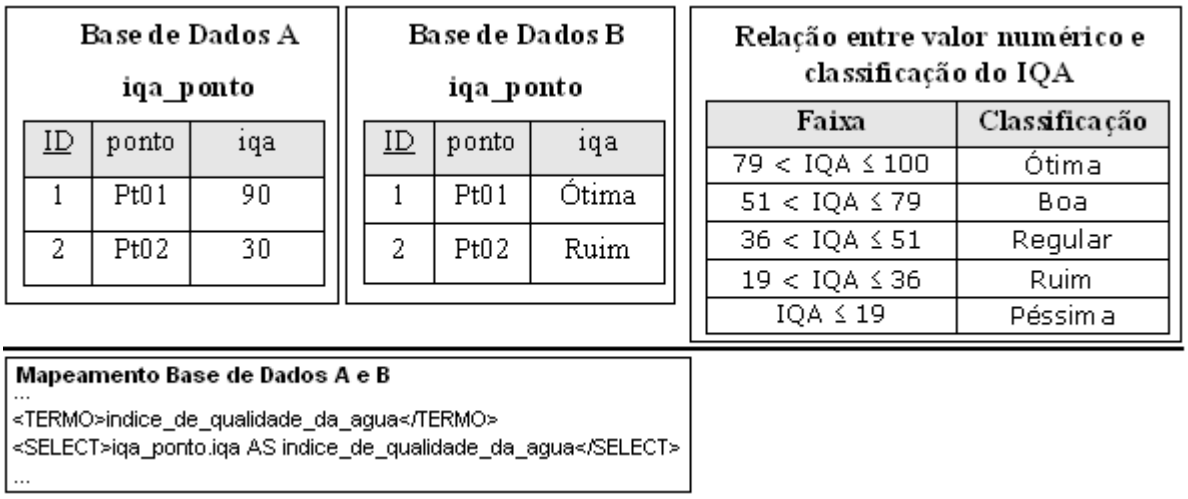


Figura 5.33 – Mapeamento para conflitos de precisão.

O outro conflito questionável é o de agregação. Um exemplo desse conflito é demonstrado na Figura 5.10. Na base de dados *A*, a entidade *peixe* armazena o comprimento total de cada indivíduo coletado e analisado. Já na base de dados *B*, cada tupla representa uma espécie de peixe analisada, contendo agregado todos os seus indivíduos analisados e uma média de seus comprimentos totais. No caso da base *A*, quando uma consulta sobre *especie_de_peixe* chega a seu tradutor, este pode traduzir a consulta à base procurando por todas as tuplas no atributo *especie*. Porém no caso da base B, a semântica para *especie_peixe* é diferente, uma vez que o tradutor vai retornar um conjunto de indivíduos representados em uma única tupla, pelos atributos *especie* e *quantidade*. Assim, esse tipo de conflito causa questões como: qual semântica deve ser utilizada pelo sistema, de individuo ou de agregação? Portanto, o mapeamento para o termo *especie_peixe* só poderá ser realizado em um dos dois casos.

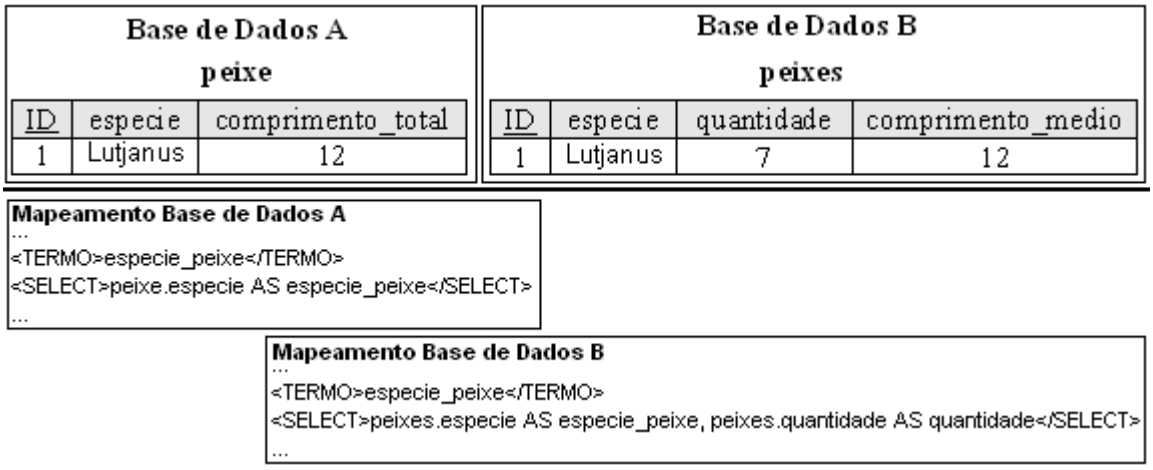


Figura 5.34 – Conflitos de agregação.

De forma geral, o DISFOQuE soluciona os desafios impostos à integração de dados, sendo a ontologia a principal responsável por tais superações.

5.6. Conclusão do capítulo

O capítulo apresentou o DISFOQuE como parte fundamental do trabalho de mestrado. Suas características gerais são de arquitetura virtual em relação às fontes e de abordagem de ontologia única em relação à forma como a ontologia é empregada no sistema. Sua arquitetura de mediador é baseada em ontologia, sendo esta a provedora do vocabulário comum utilizado no domínio. A ontologia soluciona, de forma geral, a maioria dos desafios impostos à integração de dados. Dessa forma, o sistema é flexível, aberto a fontes dispersas e autônomas e supera a maioria dos conflitos causados pelas heterogeneidades das fontes. Além disso, o sistema possui regras especiais, criadas motivadas pelas necessidades do domínio de bacias hidrográficas, porém reaproveitadas em outros domínios.

Assim, o capítulo seguinte irá aprofundar na descrição do funcionamento do DISFOQuE, focando o caso de estudo no domínio de bacias hidrográficas. Mostrará, além desse caso de estudo, outros testes de aplicação do sistema e seu desempenho.

6. O caso de estudo e outros testes

Uma vez desenvolvido o Sistema de Integração de Dados DISFOQuE, deve-se avaliar seu desempenho na consulta a dados. Principalmente no que se refere às pesquisas em fontes de dados do domínio motivador para o desenvolvimento desse sistema, que é o domínio de bacias hidrográficas com escopo nas análises das mesmas. Entretanto, outros domínios podem ser integrados pelo DISFOQuE, necessitando de uma ontologia para o domínio e dos mapeamentos entre os dados das fontes e os termos da ontologia.

Assim, este capítulo detalha o caso de estudo do domínio de bacias hidrográficas. Inicia mostrando a construção da ontologia. Depois apresenta um exemplo de uma consulta feita no DISFOQuE, integrando bancos de dados do domínio estudado. Mostra o emprego do sistema em outros domínios, citando uma outra experiência no domínio de cinema. Segue comentando o desempenho do sistema de integração e finaliza com as conclusões gerais do capítulo.

6.1. A construção da ontologia

Conforme a definição de ontologia utilizada neste trabalho, sendo "uma especificação explícita e formal de uma conceitualização compartilhada", a construção da ontologia para o domínio de bacias hidrográficas respeita cada conceito desta definição. Os conceitos, propriedades, relações, funções, restrições e axiomas, foram definidos claramente e formalmente pela linguagem OWL DL. Essas definições foram feitas sob um modelo abstrato de análises de bacias hidrográficas formado por especialistas no domínio.

A formação desse modelo abstrato é o ponto de partida para a elaboração da ontologia. Realizando estudos na área, experiência em trabalhos no IIE/IIEGA, leituras diversas, presença em palestras e seminários, conversas informais ou dirigidas sobre os vários temas que englobam o domínio, formam-se os conceitos sobre o modelo abstrato do domínio de bacias hidrográficas.

Usando como base referências como [45][56], o modelo multidisciplinar do domínio de bacias hidrográficas é elaborado e concretizado na implementação da ontologia. Ao longo desse processo houve a presença de especialistas no domínio, expressando suas concepções e formatando os resultados, o que formaliza a conceitualização compartilhada da ontologia.

A Figura 6.1 apresenta graficamente o primeiro nível de classes e suas relações constituídas na ontologia, derivadas todas da classe *"Thing"*, padrão na OWL.

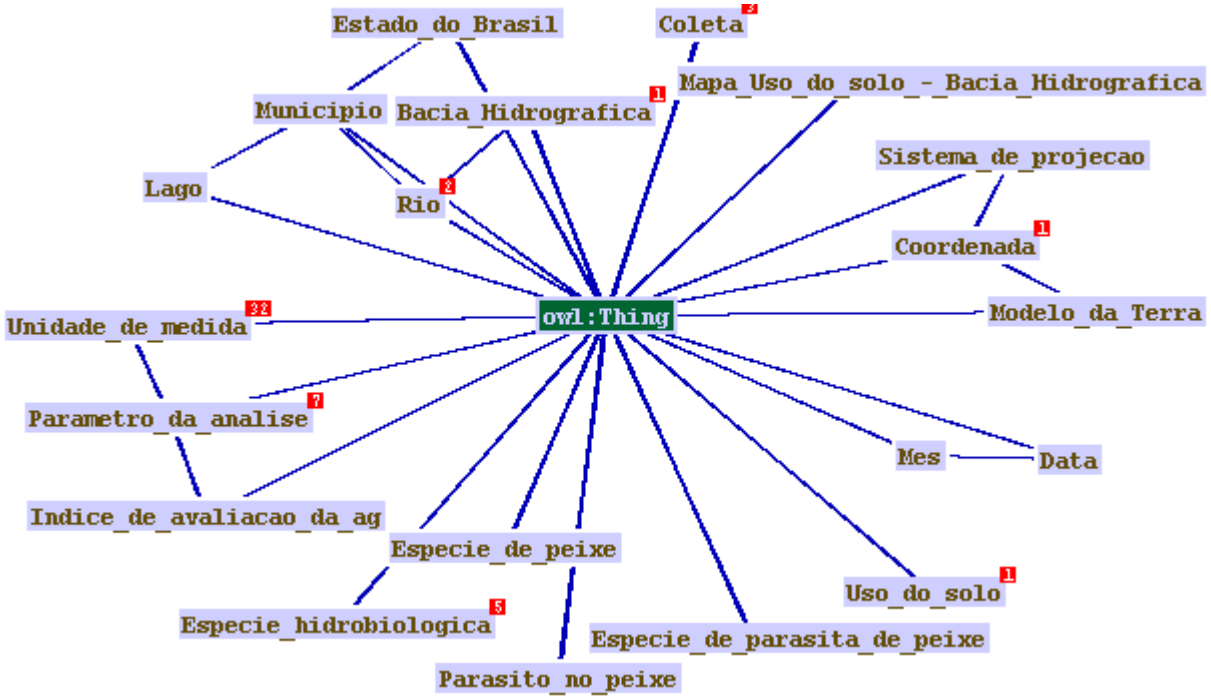


Figura 6.35 – O primeiro nível de classes da ontologia (figura gerada no software Protégé/TGVizTab).

Observa-se nessa a presença dos parâmetros relacionados ao domínio citados no capitulo 2, inseridos em suas classes. Também a questão relativa à localização no tempo e espaço está presente na ontologia. Além dessas, instâncias de bacias hidrográficas, rios e derivados, lagos, os Estados Federativos do Brasil e seus municípios, podem ser inseridas na ontologia com as relações entre elas, uma vez estabelecidas as classes e propriedades para tais atribuições. Um outro assunto pertinente tratado pela ontologia é em relação às unidades e formas de representação dos elementos formadores do domínio.

Detalhando a visão e a forma de estruturação da ontologia, a classe *Parâmetro_da_analise*, que diz respeito aos parâmetros físicos, químicos, liminologicos e aqueles que caracterizam as condições do trecho hidrográfico e do ambiente, é apresentada na Figura 6.2, tendo expandidas suas subclasses nos dois primeiros níveis de sua hierarquia.

Essa classe, suas subclasses, instâncias e propriedades, são os termos mais utilizados dentro do escopo de análise de bacias hidrográficas, ao menos dentro da maioria dos trabalhos desenvolvidos pelo IIE/IIEGA. Sua construção é baseada em documentos do instituto, como formulários de consultorias e protocolos técnicos de avaliação de água [46] [47][48], além do acompanhamento dos especialistas.

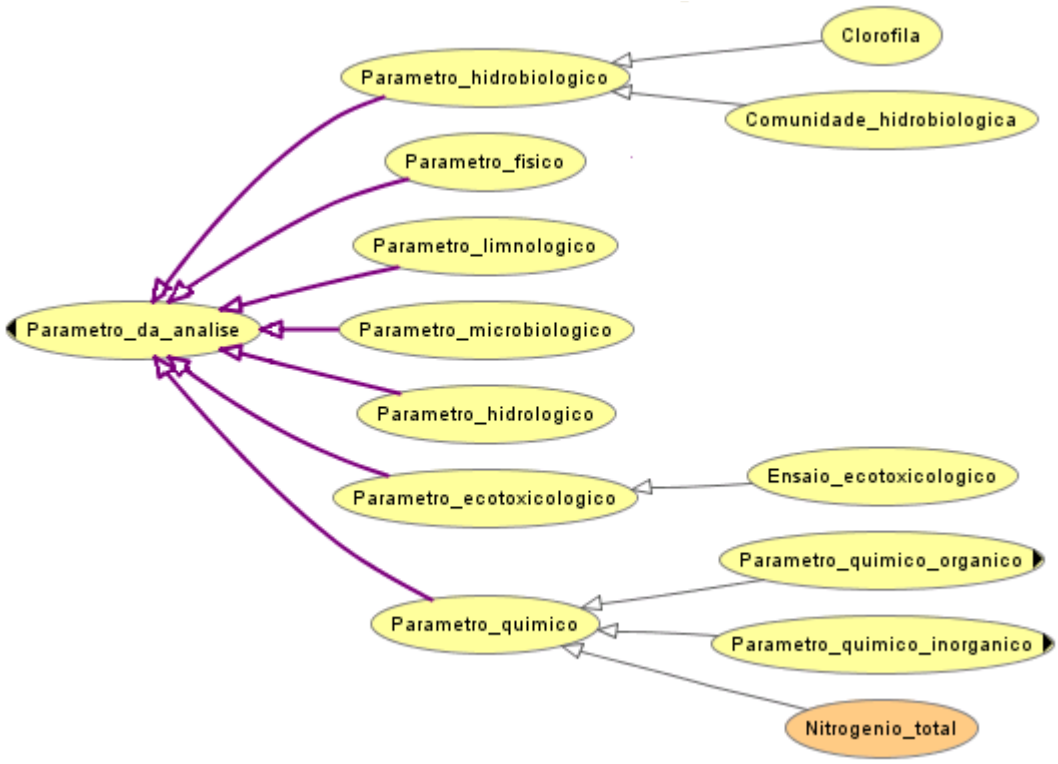


Figura 6.36 – Os dois primeiros níveis da hierarquia de sub-classes da classe *Parametro_da_analise* (figura gerada no software Protégé/OWLViz).

Aprofundando-se mais nessa classe, chegam-se às suas instâncias, que possuem propriedades que as relacionam com as unidades de medidas, os tipos de amostras e os índices de avaliação da água, como mostra um exemplo na Figura 6.3.

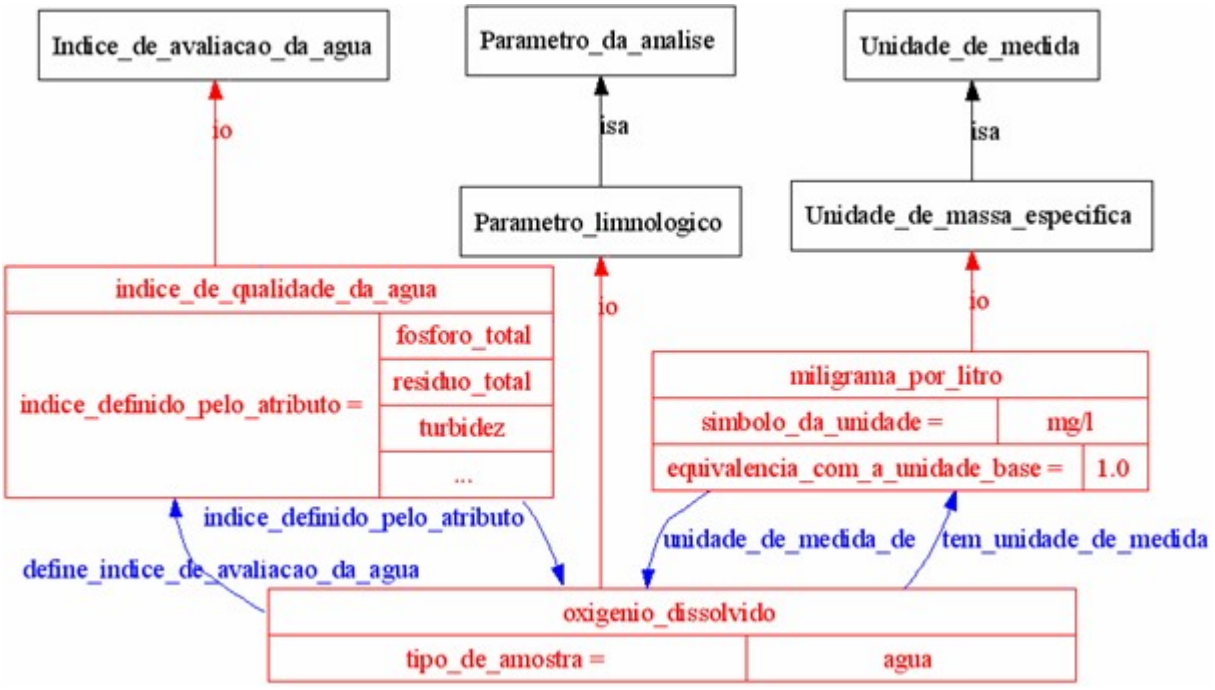


Figura 6.37 – Uma instância da classe *Parametro_da_agua*, com suas propriedades (figura gerada no software Protégé/Ontoviz).

Dessa maneira se constitui a ontologia, formada por uma hierarquia de classes, instâncias, propriedades e restrições; tudo escrito na linguagem OWL DL, formalizando a ontologia, permitindo que inferências sejam feitas por programas de computador e de forma não ambígua, deixando explícitas e formais suas especificações.

A respeito da metodologia de construção da ontologia, seguiram-se os conceitos básicos comuns às metodologias e método semelhante ao METHONTOLOGY, descritos na seção 3.4. Suas fases de desenvolvimento acompanharam o ciclo de desenvolvimento de: especificação, conceitualização, formalização e implementação, passando por revisões e manutenções. Durante todo o processo realizou-se a aquisição de conhecimento, principalmente durante as fases iniciais de criação da ontologia.

6.2. Um exemplo – da consulta à leitura dos resultados

No capítulo 5 foi apresentado o funcionamento do DISFOQuE, descrevendo seus componentes e seu fluxo de consulta (ver seção 5.2), porém, sem detalhamentos e um exemplo contínuo no fluxo da consulta. Assim, esta seção apresenta o fluxo de consulta desde a elaboração da consulta pelo usuário até a visualização dos resultados, conforme é mostrado na Figura 5.1 do capítulo 5, utilizando um exemplo no domínio de bacias hidrográficas.

O princípio é na elaboração da consulta pelo usuário. Este a faz utilizando uma interface com um campo de texto (*text field*), sem a visualização da ontologia. Para escrever a consulta algumas regras são seguidas, como utilizar as palavras *AND* e *OR* para representarem as operações lógicas E e OU respectivamente. Também as unidades de medidas devem ser colocadas entre colchetes *[]*. O Quadro 6.1 apresenta exatamente como a consulta do exemplo foi escrita pelo usuário.

OD > 2000 AND Pt_total AND temperatura AND data > 01/06/2005 AND rio = Tietê

Quadro 6.4 – Consulta escrita pelo usuário.

Confirmada a consulta, esta passa por uma análise léxica e sintática (*parser*) baseada na ontologia. Essa análise executa as seguintes ações:

- Substituição do termo *OD* por *oxigenio_dissolvido*. Ocorre devido ao termo *OD* ser um *label* (uma propriedade da RDFS) do termo *oxigenio_dissolvido* na ontologia;
- Uma vez dado um valor limite para o termo *oxigenio_dissolvido*, é verificada a unidade de medida deste valor. Não encontrada a unidade de medida, o software interage com o usuário, requisitando que ele digite uma unidade de medida para tal

valor. O usuário digita a unidade *micrograma por litro*, equivalente a 1000/1 da unidade base;

- Dada a unidade *micrograma por litro*, o valor *2000* é convertido para seu equivalente na unidade base de tal unidade de medida, ficando igual a *2*;
- O termo *Pt_total* é substituído por *fosforo_total*, como no caso do termo *OD*;
- É confirmado um caso de homônimo no termo *temperatura*, sendo este termo um *label* tanto para *temperatura_do_ar*, quanto para *temperatura_da_agua*. Nesse caso o software novamente interage com o usuário, requisitando a seleção de um dos dois termos. O usuário escolhe o termo *temperatura_do_ar*;
- O termo *data* é substituído por *Data*, como no caso do termo *OD*;
- A formatação da data é analisada e esta fica sendo formatada como: *20050601*;
- O termo *rio* é substituído por *Rio*, como no caso do termo *OD*.

Uma vez executadas essas ações a consulta fica como mostrada no Quadro 6.2.

oxigenio_dissolvido > 2 AND fosforo_total AND temperatura_do_ar AND Data > 20050601 AND Rio = Tietê

Quadro 6.5 - Consulta após ações executadas pelo *Parser*.

Após as ações executadas pelo *Parser* da consulta, esta é enviada ao módulo do sistema FOQuE, onde será expandida conforme regras predeterminadas expressas na ontologia. Dessa forma, o termo *fosforo_total* é expandido para *nitrogenio_total*, uma vez que na ontologia existe uma similaridade semântica entre estes dois termos. Essa similaridade possui um valor difuso de *0.8*, ou seja, 80%, sendo maior que os *0.7* da similaridade mínima de corte, configurada no sistema pelo usuário. O Quadro 6.3 apresenta como a consulta ficou após a passagem pelo módulo do sistema FOQuE.

oxigenio_dissolvido > 2 AND fosforo_total OR nitrogenio_total AND temperatura_do_ar AND Data > 20050601 AND Rio = Tietê

Quadro 6.6 – Consulta após a expansão pelo módulo do sistema FOQuE.

Expandida a consulta, esta é enviada ao mediador do sistema, que a distribui, exatamente como ela chegou, aos tradutores cadastrados. O arquivo de cadastro dos tradutores desse exemplo é o mostrado no Quadro 6.4.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE root [
        <!ELEMENT root ( WRAPPER+ ) >
        <!ELEMENT WRAPPER ( ID, NOME, CONEXAO ) >
        <!ELEMENT ID ( #PCDATA ) >
        <!ELEMENT NOME ( #PCDATA ) >
        <!ELEMENT CONEXAO ( ENDERECO_IP, PORTA )>
        <!ELEMENT ENDERECO_IP ( #PCDATA ) >
        <!ELEMENT PORTA ( #PCDATA ) >
        <!ATTLIST root version NMTOKEN #REQUIRED >
]>
<root>
        <WRAPPER>
                <ID>1</ID>
                <NOME>Tradutor do BD do projeto de politicas publicas</NOME>
                <CONEXAO>
                        <ENDERECO_IP>127.0.0.1</ENDERECO_IP>
                        <PORTA>50001</PORTA>
                </CONEXAO>
        </WRAPPER>
        <WRAPPER>
                <ID>2</ID>
                <NOME>Tradutor do BD com dados hidrologicos</NOME>
                <CONEXAO>
                        <ENDERECO_IP>192.168.0.5</ENDERECO_IP>
                        <PORTA>50002</PORTA>
                </CONEXAO>
        </WRAPPER>
</root>
```

Quadro 6.7 – Arquivo com o cadastro dos tradutores.

Estabelecida uma conexão com cada tradutor descrito no arquivo de cadastro, o mediador repassa a consulta aos tradutores. A consulta então será adaptada por cada tradutor a sua fonte de dados, por meio do arquivo de mapeamento que cada fonte possui em seu tradutor particular. Caso o tradutor não possua nenhum termo referente à consulta mapeado para sua fonte de dados, ele encerra sua atividade naquela consulta, fechando sua conexão com o mediador sem retornar dado algum. Todavia, se houver algum termo na consulta, exceto termos chaves na fonte, como geralmente as datas e coordenadas, é feita uma consulta para cada um dos termos, conforme seu mapeamento no arquivo do tradutor. As consultas realizadas na fonte são integradas à medida que são realizadas, sendo respeitadas as precedências nas operações e utilizando os preceitos da álgebra booleana, realizando para a lógica E (*AND*) a intersecção dos resultados e para a lógica OU (*OR*) a união dos resultados. Nesse exemplo, a Figura 6.4 apresenta o modelo e os dados das duas fontes consultadas pelo DISFOQuE. Ambas as fontes estão no modelo relacional, tendo como linguagem de consulta a SQL.

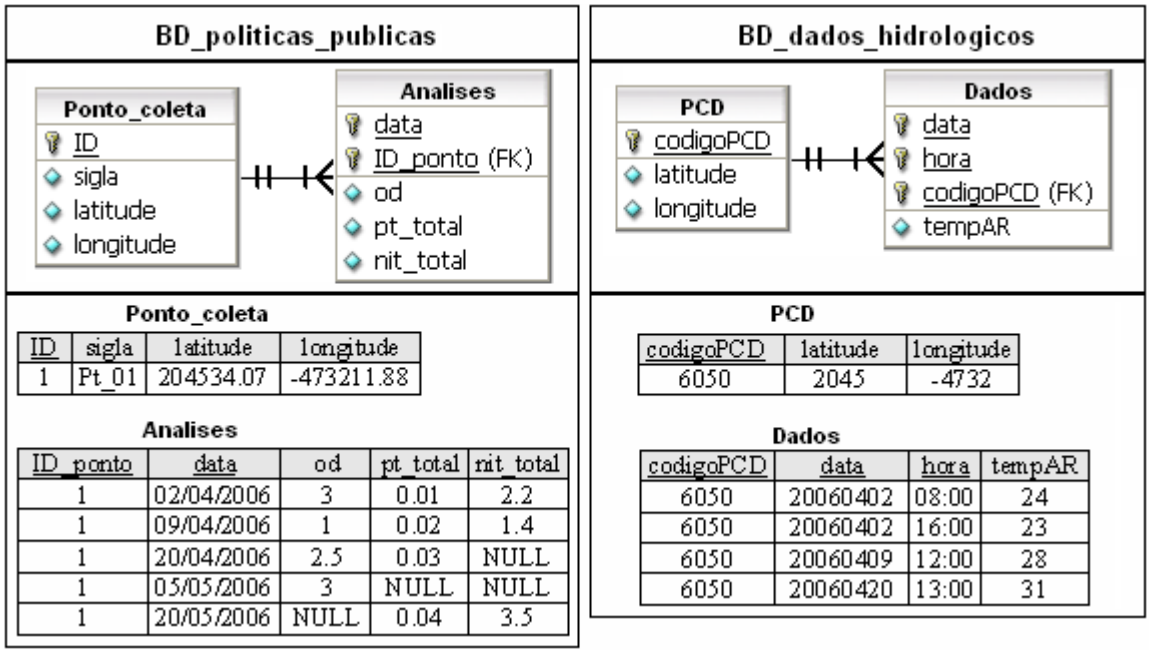


Figura 6.38 – Bancos de Dados dos tradutores do DISFOQuE.

Uma vez que a consulta chega aos tradutores das fontes, cada um deles fará a respectiva tradução da consulta (Quadro 6.3) para sua fonte e executará as rotinas necessárias, conforme descrições a seguir. Será narrada em detalhes apenas a execução dessa tarefa na fonte *BD_politicas_publicas*, sendo que para a fonte *BD_dados_hidrologicos* a execução ocorre de forma similar.

O arquivo de mapeamento do tradutor de *BD_politicas_publicas* é dado no Quadro 6.5 seguinte.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE root [
<!ELEMENT root ( SOCKET ) >
<!ELEMENT root ( CONEXAO_BD_LOCAL ) >
<!ELEMENT CONEXAO_BD_LOCAL ( SGBD, BD_NOME, ENDERECO_IP, PORTA, USUARIO, SENHA ) >
<!ELEMENT root ( CONEXAO_BD_GEOGRAFICO ) >
<!ELEMENT CONEXAO_BD_GEOGRAFICO ( SGBD, BD_NOME, ENDERECO_IP, PORTA, USUARIO, SENHA ) >
<!ELEMENT root ( PADROES_FORMATOS ) >
<!ELEMENT PADROES_FORMATOS (FORMA_COORDENADAS, FORMA_DATA) >
<!ELEMENT root ( MAPEAMENTO+ ) >
<!ELEMENT MAPEAMENTO ( TERMO, SELECT*, FROM*, WHERE*, WHERE_PARAMETRO*,
TIPO_PROCESSAMENTO*, FATOR_TRANSFORMA_UNIDADE*, CHAVE*) >
<!ELEMENT TERMO ( #PCDATA ) >
<!ELEMENT SELECT ( #PCDATA ) >
<!ELEMENT FROM ( #PCDATA )>
<!ELEMENT WHERE ( #PCDATA ) >
<!ELEMENT WHERE_PARAMETRO ( #PCDATA ) >
<!ELEMENT TIPO_PROCESSAMENTO ( #PCDATA ) >
<!ELEMENT FATOR_TRANSFORMA_UNIDADE ( #PCDATA ) >
<!ELEMENT FORMA_COORDENADAS ( #PCDATA ) >
<!ELEMENT CHAVE ( #PCDATA ) >
```

```
<!ATTLIST root version NMTOKEN #REQUIRED >
]>
<root>
	<SOCKET>50001</SOCKET>
	<CONEXAO_BD_LOCAL>
		<SGBD>mysql</SGBD>
		<BD_NOME>BD_politicas_publicas</BD_NOME>
		<ENDERECO_IP>localhost</ENDERECO_IP>
		<PORTA>3306</PORTA>
		<USUARIO>sid</USUARIO>
		<SENHA>123</SENHA>
	</CONEXAO_BD_LOCAL>
	<CONEXAO_BD_GEOGRAFICO>
		<SGBD>postgresql</SGBD>
		<BD_NOME>bd_geografico</BD_NOME>
		<ENDERECO_IP>localhost</ENDERECO_IP>
		<PORTA>5432</PORTA>
		<USUARIO>postgres</USUARIO>
		<SENHA>123456</SENHA>
	</CONEXAO_BD_GEOGRAFICO>

	<PADROES_FORMATOS>
		<FORMA_COORDENADAS>GMS</FORMA_COORDENADAS>
		<FORMA_DATA>DD/MM/AAAA</FORMA_DATA>
	</PADROES_FORMATOS>

	<MAPEAMENTO>
		<TERMO>oxigenio_dissolvido</TERMO>
		<SELECT>Analises.od AS oxigenio_dissolvido</SELECT>
		<FROM>Analises, Ponto_coleta</FROM>
		<WHERE>Analises.od IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID</WHERE>
		<WHERE_PARAMETRO>Analises.od $sinal $valor</WHERE_PARAMETRO>
		<CHAVE>Ponto_coleta.latitude AS chave_latitude</CHAVE>
		<CHAVE>Ponto_coleta.longitude AS chave_longitude</CHAVE>
		<CHAVE>Analises.data AS chave_Data</CHAVE>
	</MAPEAMENTO>
	<MAPEAMENTO>
		<TERMO>fosforo_total</TERMO>
		<SELECT>Analises.pt_total AS fosforo_total</SELECT>
		<FROM>Analises, Ponto_coleta</FROM>
		<WHERE>Analises.pt_total IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID</WHERE>
		<WHERE_PARAMETRO>Analises.pt_total $sinal $valor</WHERE_PARAMETRO>
		<CHAVE>Ponto_coleta.latitude AS chave_latitude</CHAVE>
		<CHAVE>Ponto_coleta.longitude AS chave_longitude</CHAVE>
		<CHAVE>Analises.data AS chave_Data</CHAVE>
	</MAPEAMENTO>
	<MAPEAMENTO>
		<TERMO>nitrogênio_total</TERMO>
		<SELECT>Analises.nit_total AS nitrogênio_total</SELECT>
		<FROM>Analises, Ponto_coleta</FROM>
		<WHERE>Analises.nit_total IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID</WHERE>
		<WHERE_PARAMETRO>Analises.nit_total $sinal $valor</WHERE_PARAMETRO>
		<CHAVE>Ponto_coleta.latitude AS chave_latitude</CHAVE>
		<CHAVE>Ponto_coleta.longitude AS chave_longitude</CHAVE>
		<CHAVE>Analises.data AS chave_Data</CHAVE>
	</MAPEAMENTO>

	<MAPEAMENTO>
		<TERMO>Rio</TERMO>
		<SELECT></SELECT>
```

```
                <FROM>Ponto_coleta</FROM>
                <TIPO_PROCESSAMENTO>GEOGRAFICO</TIPO_PROCESSAMENTO>
                <CHAVE>Ponto_coleta.latitude AS chave_latitude</CHAVE>
                <CHAVE>Ponto_coleta.longitude AS chave_longitude</CHAVE>
        </MAPEAMENTO>
        <MAPEAMENTO>
                <TERMO>Data</TERMO>
                <TIPO_PROCESSAMENTO>CHAVE</TIPO_PROCESSAMENTO>
                <SELECT>chave_Data</SELECT>
                <WHERE_PARAMETRO>chave_Data $sinal $valor</WHERE_PARAMETRO>
        </MAPEAMENTO>
</root>
```

Quadro 6.8 – Arquivo de mapeamento do tradutor da fonte *BD_politicas_publicas.*

Conforme o arquivo do Quadro 6.5, a consulta do Quadro 6.3 é realizada no tradutor de *BD_politicas_publicas* conforme o Quadro 6.6 seguinte.

1. *SELECT Analises.latitude AS chave_latitude, Analises.longitude AS chave_longitude, Analises.data AS chave_Data, Analises.pt_total AS fosforo_total FROM Analises, Ponto_coleta WHERE Analises.pt_total IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID AND Analises.data > 01/06/2005*

Faz a união desses resultados com os resultados da consulta seguinte (U)

2. *SELECT Analises.latitude AS chave_latitude, Analises.longitude AS chave_longitude, Analises.data AS chave_Data, Analises.nit_total AS nitrogenio_total FROM Analises, Ponto_coleta WHERE Analises.nit_total IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID AND Analises.data > 01/06/2005*

Faz a intersecção dos resultados da união das duas primeiras consultas com os resultados da consulta seguinte (∩)

3. *SELECT Analises.latitude AS chave_latitude, Analises.longitude AS chave_longitude, Analises.data AS chave_Data, Analises.od AS oxigenio_dissolvido FROM Analises, Ponto_coleta WHERE Analises.od IS NOT NULL AND Analises.ID_ponto = Ponto_coleta.ID AND Analises.data > 01/06/2005*

Quadro 6.9 – Consultas realizadas na fonte *BD_politicas_publicas*.

Após essas consultas feitas na fonte *BD_politicas_publicas*, os resultados retornados passam por um processamento consultando o BD Geográfico, uma vez que o termo *Rio* necessita deste processamento. Assim, as coordenadas dos dados retornados do resultado das consultas do Quadro 6.6 são enviadas a esse processamento geográfico, que retorna apenas as coordenadas que fazem parte do rio *Tietê*. Nesse exemplo todos os dados em ambos os Bancos têm suas coordenadas pertencentes ao rio Tietê.

Terminado o processamento, os resultados que o tradutor da fonte *BD_politicas_publicas* retorna ao mediador são apresentados na Tabela 6.1. Na Tabela 6.2 são apresentados os resultados retornados ao mediador pelo tradutor da fonte *BD_dados_hidrologicos*, que executou de forma similar as tarefas do tradutor *BD_politicas_publicas*, descritas anteriormente.

chave_latitude	chave_logitude	chave_Data	oxigenio_dissolvido	fosforo_total	nitrogenio_total
204534.07	-473211.88	20060402	3	0.01	2.2
204534.07	-473211.88	20060420	2.5	0.03	NULL

Tabela 6.5 – Resultados retornados da fonte *BD_politacas_pubicas*.

chave_latitude	chave_logitude	chave_Data	chave_hora	Temperatura_do_ar
2045	-4732	20060402	08:00	24
2045	-4732	20060402	16:00	23
2045	-4732	20060404	12:00	28
2045	-4732	20060420	13:00	31

Tabela 6.6 – Resultados retornados da fonte *BD_dados_hidrologicos*.

Uma vez os resultados no mediador, este os integra, utilizando para tal as mesmas premissas da álgebra booleana empregada pelos tradutores, ou seja, a lógica E (*AND*) realizando as intersecções entre os resultados e a lógica OU (*OR*) realizando as uniões entre os resultados. Dessa forma, os dados retornados pelos tradutores ao mediador são integrados baseando-se na consulta do Quadro 6.3, concluindo-se na Tabela 6.3 apresentada a seguir.

chave_latitude	chave_logitude	chave_Data	chave_hora	oxigenio_dissolvido	fosforo_total	nitrogenio_total	temperatura_do_ar	Rio
204534.07	-473211.88	20060402	08:00	3	0.01	2.2	24	Tiete
204534.07	-473211.88	20060402	16:00	3	0.01	2.2	23	Tiete
204534.07	-473211.88	20060420	13:00	2.5	0.03	NULL	31	Tiete

Tabela 6.7 – Integração dos dados consultados nas fontes do DISFOQuE.

Terminada a integração dos resultados, os dados são retornados à camada de formulação da consulta e apresentação dos resultados, onde são exibidos ao usuário, formatados em uma tabela semelhante à Tabela 6.3, concluindo o ciclo da consulta e integração de dados.

6.3. O emprego do sistema de integração em outros domínios

Como enunciado anteriormente, o DISFOQuE pode ser utilizado para integrar dados de outro domínio e não somente de seu domínio inspirador. É certo que algumas características desse sistema foram criadas especialmente ao domínio de bacias hidrográficas,

como o BD Geográfico, e assim são utilizadas quase que exclusivamente neste domínio. Porém, sua função principal de integração de dados é desempenhada em qualquer domínio que possua uma ontologia OWL, fontes e seus respectivos tradutores. Assim, o DISFOQuE torna-se bastante flexível, permitindo a integração de fontes de domínios diversos.

Para testar essa flexibilidade, foi utilizado um domínio de cinema, citado em referências como em [49]. Dessa forma, utilizando-se esse domínio, nenhuma característica especial do DISFOQuE será empregada, como o tratamento de unidades de medidas e dados espaciais. Apenas suas funções básicas de integração de dados são realizadas, tendo-se assim um parâmetro de comparação no desempenho do sistema entre domínios que fazem uso de processamentos especiais e os que não fazem, permitindo uma estimativa do gasto com tais processamentos especiais, principalmente com o processamento geográfico.

No domínio de cinema três Bancos de Dados Relacionais foram criados e povoados com dados fictícios de tal domínio. A primeira fonte contém os dados dos filmes com o título original, o título em português, o ano de lançamento, os gêneros e o diretor. Outra fonte armazena dados sobre as premiações dadas aos filmes, tendo os atributos de nome do filme, o premio recebido e o ano em que recebeu. Uma terceira fonte relaciona as críticas dadas aos filmes, com o nome original do filme, o nome em português, a crítica e o crítico.

Uma ontologia foi desenvolvida contendo os termos básicos desse domínio. A Figura 6.5 apresenta graficamente essa ontologia, sem muitos detalhes.

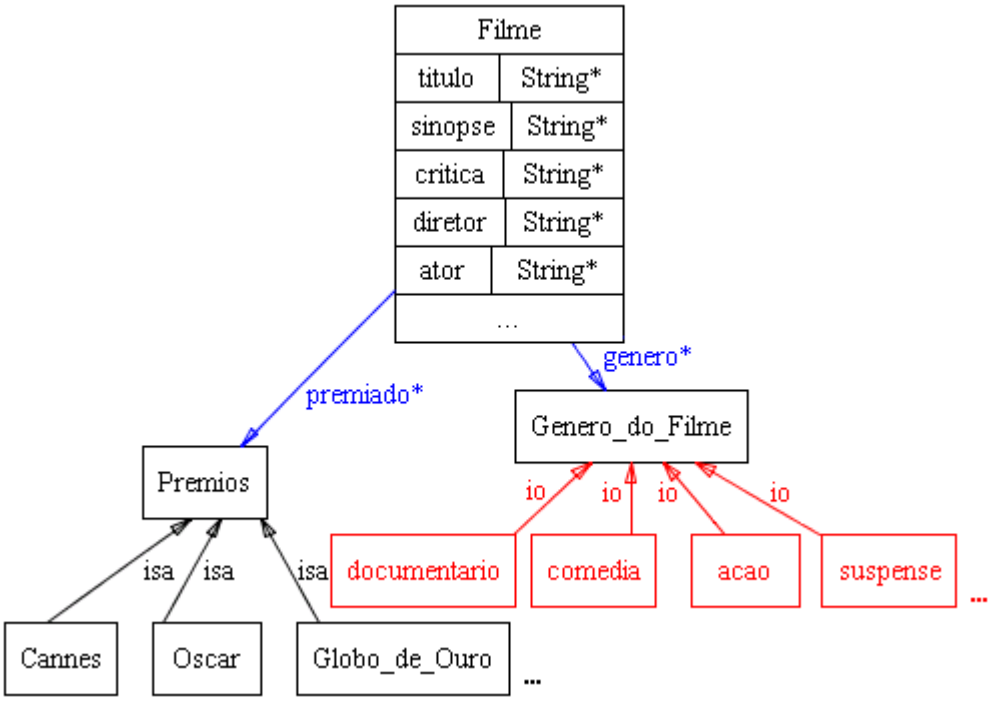


Figura 6.39 – Ontologia do domínio de cinema.

Com essa ontologia é possível consultar o domínio de cinema, pesquisando as três fontes disponíveis. Assim, consultas como: as críticas aos filmes do diretor Steven Spielberg que ganharam Oscar, são passíveis a resposta do sistema, uma vez que este consegue consultar em todas as fontes e integrar seus dados, relacionando-os. Esse exemplo de consulta é bastante didático na explicação da função desempenhada pelos sistemas de integração, pois permite visualizar claramente que a consulta não poderia ser respondida se fossem feitas buscas isoladas a cada uma das fontes de dados, necessitando de uma integração entre os dados das fontes, como se elas fossem uma única fonte, formada pela visão global do sistema.

6.4. Desempenho do sistema de integração

Feitos os testes com o DISFOQuE, tanto no domínio principal, de bacias hidrográficas, quanto em outro domínio qualquer, de cinema, é feito nesta seção uma análise do desempenho deste sistema de integração. Uma investigação de quais são os gargalos pertinentes à integração de dados nos tempos de processamento para a busca e integração dos dados. Os tempos dos testes mencionados neste documento foram levantados em uma máquina com as seguintes especificações: processador Athlon64 X2 4200, 2 GB de memória RAM e Sistema Operacional Windows XP. A respeito dos Bancos de Dados utilizados, todos os testes foram feitos com Bancos Relacionais gerenciados pelo SGBD MySQL 5.0.41, com exceção do fonte *BD_cptec_inpe* (arquivo XML), e todos os Bancos são armazenados localmente na máquina. Dessa forma pretende-se medir o desempenho do processamento dos dados referentes à sua integração, isolando fatores como a busca interna na base de dados e tráfego na rede. A constituição dos Bancos de Dados utilizados é mostrada na Tabela 6.4 a seguir.

Banco de Dados	Entidade	Atributos	Quantidade de registros
BD_politicas_publicas	ponto	ident, nome, latitude, longitude, altitude	44
	analises	ident_ponto, data, profundidade, oxi_dis, ph, dbo, fos_total, nit_total	86
BD_brasil_das_aguas	ponto	id, nome, latitude, lontitude	1157
	dados	id, od, ph, ptotal	1157
*BD_cptec_inpe (*XML)	pcd	codigo, lat, long, alt	1
	dados	codigoPCD, dataHora, tempAR	11251
BD_hidrica	ponto	id, nome, lat, long,alt	4
	característica_hidrica	id_ponto, data, vazao, precipitacao	15546
BD_cinema_filme	filme	id, titulo_original, titulo_portugues, ano, genero, diretor	10
BD_cinema_premiacao	premiação_filme	id, nome_filme, premio, ano	1000
BD_cinema_critica	criticas_filme	id, nome_original, nome_nacional, critica, nome_critico	10000

Tabela 6.8 – Bancos de Dados Relacionais utilizados nos testes.

Seguindo o fluxo de uma consulta, o primeiro passo é a formulação da consulta que será distribuída às fontes do sistema. Essa etapa, que vai desde a formulação da consulta pelo usuário até a chegada da mesma ao mediador do sistema, não é um gargalo para o sistema. Seus tempos de processamentos são bastante curtos, sendo que a maior parte do tempo é na carga da ontologia difusa e nas regras referentes ao módulo do sistema FOQuE. Mesmo com esses processamentos extras, o tempo não ultrapassa dois segundos, não sofrendo grandes variações, independente da complexidade da consulta a ser realizada.

A etapa seguinte mede o desempenho a partir da distribuição da consulta aos tradutores do sistema até o retorno dos resultados das consultas ao mediador. Nessa etapa o processamento árduo é desempenhado pelos tradutores do sistema na realização das consultas e processos de relacionamento de seus dados, realizados pelas uniões e intersecções destes. Além desses processos padrões, muitas consultas necessitam de processamentos especiais, como consultas ao BD Geográfico, que também consome intensamente o tempo total de processamento. Dessa forma, o tempo de processamento sofre um alto grau de variação dependendo da complexidade da consulta. Consultas simples, que não exigem nenhum processamento especial e exige pouco ou nenhum cruzamento entre os dados pelo tradutor, são realizadas em tempos pequenos, por volta de um segundo. Já as consultas mais

complexas, que envolvem grandes cruzamentos entre os dados, realizando as uniões e intersecções entre eles, consomem tempos relativamente elevados, como algumas dezenas de segundos, podendo chegar a minutos, dependendo da complexidade da consulta, do número de registros pesquisados e do desempenho da máquina que roda o processo. Os processamentos especiais também consomem grande parte do tempo total do processamento, casos entre 20% a 30% do tempo total de processamento da consulta.

A última etapa inicia-se na integração dos dados retornados dos tradutores ao mediador e finaliza-se com apresentação desses dados ao usuário. O gargalo nessa etapa está na integração dos dados feita pelo mediador, uma vez que estes integrados, sua formatação para apresentação é simples, portanto rápida.

No mediador todos os dados que são retornados dos tradutores são integrados, relacionando-os entre si pelos atributos chaves (chaves naturais), assim como foi feito em cada um dos tradutores, com a união e intersecção dos dados. Porém, a integração dos dados no mediador é mais custosa, uma vez que uma quantidade maior de dados será trabalhada e uma quantidade maior de cruzamentos entre os registros deverá ser realizada. Além do mais, relacionamentos de semelhança entre valores de chaves iguais são testados em alguns casos. Por exemplo, quando a *chave_Data* está presente em mais de uma fonte que retornou dados para a consulta, os relacionamentos feitos entre estes dados verifica se a quantidade de caracteres de um registro é diferente de outro, pois sendo, apenas parte da cadeia de caracteres do registro maior (quantidade maior de caracteres no dado) será comparada. Como no caso da data *200604* em um registro de uma fonte e *20060411* em um registro de outra fonte, serão comparados apenas os seis primeiros caracteres do registro da ultima fonte, concluindo-se que são datas semelhantes. Esse tipo de comparação requer uma quantidade de processamento maior, levando mais tempo para ser realizada.

Mesmo com essas dificuldades maiores na integração de dados, o tempo de processamento dessa etapa pode ser inferior ao tempo da etapa anterior, pois os dados processados já foram filtrados na etapa anterior, o que consome grande parte do tempo total do ciclo da consulta. Entretanto, se uma quantidade grande de registros de fontes diferentes (acima de 2000 registros por fonte, bastando duas fontes) for enviada ao mediador, esse tempo de integração dos dados passa a ser elevado, superando o tempo gasto nas consultas às fontes. Os tempos de processamento dessa etapa também são bastante variáveis, dependendo da quantidade de registros, do tipo de chaves que estes possuem e da complexidade lógica da consulta. Em testes com consultas mais simples, esse tempo fica em torno de um a três segundos, mesmo com uma quantidade de registros razoável. Em consultas mais complexas

que exigem um número maior de relacionamento entre as chaves, o tempo de processamento dessa etapa aumenta cerca de três segundos comparado com uma consulta mais simples que retorna uma quantidade de registro equivalente.

A Tabela 6.5 apresenta o resultado do desempenho de algumas consultas feitas no DISFOQuE, utilizando-se a máquina e os Bancos de Dados descritos no início desta seção.

Consulta	Fontes consultadas	Quantidade de registros retornados		Tempo total da consulta [segundos]
		Por fonte	Após a integração	
oxigênio dissolvido	BD_politicas_publicas	85	1232	11
	BD_brasil_das_aguas	1147		
oxigênio dissolvido AND dbo	BD_politicas_publicas	55	55	11
	BD_brasil_das_aguas	1147		
oxigênio dissolvido > 3	BD_politicas_publicas	85	1072	11
	BD_brasil_das_aguas	987		
oxigênio dissolvido AND bacia = Feijao	BD_politicas_publicas	85	48	13
	BD_brasil_das_aguas	1147		
oxigênio dissolvido AND temperatura do ar	BD_politicas_publicas	85	433	76
	BD_brasil_das_aguas	1147		
	BD_cptec_inpe	11187		
oxigênio dissolvido AND temperatura do ar AND bacia = Feijao	BD_politicas_publicas	85	373	77
	BD_brasil_das_aguas	1147		
	BD_cptec_inpe	11187		
precipitação	BD_hidrica	13803	13803	77
precipitação AND temperatura do ar	BD_hidrica	13803	5450	347
	BD_cptec_inpe	11187		
críticas	BD_cinema_criticas	9426	9426	20
críticas AND diretor = Spielberg AND premiação = Oscar	BD_cinema_criticas	9426	11372	23
	BD_cinema_premiacao	40		
	BD_cinema_filme	3		
gênero = suspense OR gênero = terror AND premiação	BD_cinema_premiacao	100	20	3
	BD_cinema_filme	2		

Tabela 6.9 – Desempenho de algumas consultas realizadas no SID.

6.5. Conclusão do capítulo

O DISFOQuE foi idealizado pensando-se na integração de dados de análises de bacias hidrográficas. Dessa forma, criaram-se funções especiais para esse domínio, como o processamento de dados georreferenciados e tratamento de unidades de medidas. Parte considerável desse trabalho foi o desenvolvimento da ontologia para esse domínio,

possibilitando uma visão global de seus termos, classificados em uma hierarquia de classes e relacionando-se entre si com propriedades diversas. Entretanto, o DISFOQuE não atende somente a esse domínio específico, sendo flexível o suficiente para se adaptar a outros domínios, bastando para tal a criação de uma ontologia em OWL e do mapeamento entre os termos desta ontologia com os termos das fontes.

O desempenho do sistema em relação ao tempo de processamento é um fator crítico. O tempo de resposta às consultas é variável, dependendo da complexidade da consulta, dos processamentos especiais realizados e principalmente da quantidade de registros consultados. É necessário que o sistema rode em máquinas com alto desempenho de processamento, uma vez não satisfeito tal requisito o sistema pode ter tempos de resposta elevados, a ponto de inviabilizar seu uso.

Após a descrição da função e detalhado o funcionamento do DISFOQuE, o capítulo seguinte conclui este documento com uma discussão sobre os trabalhos relacionados e os resultados atingidos, além de apontar trabalhos futuros.

7. Conclusões gerais

Como base para o desenvolvimento deste trabalho, fez-se o levantamento bibliográfico, pesquisando sobre o tema ontologia e Sistema de Integração de Dados, de forma isolada e em conjunto, ou seja, SIDs baseados em ontologias. Também, com igual importância para o fundamento do trabalho está a experiência adquirida nas pesquisas desenvolvidas no IIE, que motivaram o desenvolvimento deste trabalho, além de estudos e práticas com SIGs e Banco de Dados Geográficos, necessários em tal domínio.

A partir desses pilares, deu-se o desenvolvimento do DISFOQuE, apresentado como elemento central do trabalho de mestrado. Esse sistema, que tem a ontologia como base, permitindo a visão única e transparente às consultas nas fontes, cumpre as funções inerentes aos SIDs, além de possuir funções especiais, como a expansão de consulta e tratamento de dados geográficos. Apesar de sua visualização ter surgido sobre o domínio de análises de bacias hidrográficas, realizadas no IIE, sua aplicação é expansível a outros domínios.

Assim, após a descrição do domínio de bacias hidrográficas, do levantamento bibliográfico, que inclui ontologia e Sistemas de Integração de Dados, do DISFOQuE e de testes e análise de desempenho deste sistema, faz-se neste capítulo as conclusões gerais referentes ao trabalho aqui apresentado. Inicialmente são discutidos os trabalhos relacionados e os resultados atingidos. Finaliza-se com as propostas de trabalhos futuros.

7.1. Trabalhos relacionados e resultados atingidos

O tema integração de dados não é recente, tendo sido abordado e identificado como criticamente importante desde a criação dos Sistemas Gerenciadores de Dados [50], destacando-se em meio à necessidade de buscas de informações em um ambiente de crescente disponibilidade de dados em fontes distribuídas e heterogêneas. Assim, o desenvolvimento de Sistemas de Integração de Dados, que interagem essas fontes provendo uma visão única e de forma transparente ao usuário, vem seguindo em paralelo a essa crescente necessidade de integrar dados.

A primeira distinção desses sistemas feita nesta seção é em relação à presença de uma ou conjunto de ontologias. O beneficio no uso de ontologias está no tratamento semântico dos dados. Esta característica é empregada principalmente nas soluções aos desafios da heterogeneidade lógica (ver seção 4.2.4 e 5.5) das fontes envolvidas no sistema.

Com base nessa distinção entre os SIDs, a seção 4.5 apresenta alguns dos principais SIDs baseados em ontologia encontrados na literatura sobre tal tema. Em comparação a esses sistemas, o DISFOQuE assemelha-se aos sistemas SIMS e TAMBIS, uma vez que as abordagens de uso de uma ontologia única e arquitetura mediador/tradutor são comuns. A semelhança com o sistema TAMBIS é visível analisando-se as Figuras 4.16 (do sistema TAMBIS) e 5.1 (do DISFOQuE) e mesmo em relação a ambos serem destinados a um domínio especifico, sendo o TAMBIS destinado ao domínio de bioinformática e o DISFOQuE destinado especialmente ao domínio de bacias hidrográficas. Entretanto é válido ressaltar novamente que o DISFOQuE é flexível para trabalhar com outros domínios de temas diversos. Apesar de seu desenvolvimento ter sido motivado para integrar dados de análises de bacias hidrográficas, o que gerou funcionalidades especiais para tal, esse sistema pode ser utilizado em qualquer domínio que possua uma ontologia OWL desenvolvida e fontes estruturadas em Bancos de Dados Relacionais ou Orientado a Objetos, ou ainda semi-estruturadas em XML, igualmente a outros sistemas, como o sistema Integra [30].

O que se pode destacar de especial no DISFOQuE são as funcionalidades desenvolvidas motivadas pelo domínio de aplicação principal, de bacias hidrográficas. Essas funcionalidades próprias são:

- Tratamento das unidades de medidas. A ontologia construída para o domínio aborda semanticamente as relações entre as unidades de medidas, classificando-as e relacionando-as com valores de transformação entre as unidades da mesma grandeza;
- Comparações entre coordenadas, verificando se um valor de coordenada está incluso em outro valor de diferente fonte. Essa comparação é realizada checando-se o tamanho da cadeia de caracteres (*string*) das coordenadas, igualando a comparação ao menor comprimento de cadeia. Da mesma forma, essa função é realizada entre os atributos de identificação temporal (datas e horas);
- Banco de Dados Geográfico, auxiliando a identificação, e assim integração, dos dados. É comum na literatura sobre integração de dados o tema de integração de dados geográficos, buscando uma forma comum de representação destes dados. Entretanto, o enfoque aos dados geográficos no sistema desenvolvido neste trabalho é outro. Aqui há um único BD Geográfico, com dados referentes ao domínio de bacias hidrográficas (rios, limites de bacias hidrográficas, limites de estados da federação brasileira, relação entre os rios e as bacias hidrográficas e entre as próprias bacias etc), que tem a função de fornecer informações sobre a localização de pontos de coordenadas

enviadas a ele, além de informar a ontologia sobre as bacias hidrográficas e rios existentes e os relacionamentos entre estes objetos.

Além dessas funcionalidades especiais, há a expansão da consulta obtida pelo módulo do sistema FOQuE, que permite a expansão, segundo conceitos da lógica difusa, pelos seguintes tipos:

- Similaridade: possibilita recuperar conceitos semanticamente relevantes considerando graus de similaridades entre conceitos. Para tal é definido na ontologia pelos especialistas, um grau de similaridade entre as instâncias consideradas similares. Por exemplo, as instâncias *demanda química de oxigenio* e *demanda bioquímica de oxigenio* são consideradas similares com um grau de *0.75*. Assim, em uma consulta por *demanda química de oxigenio*, considerando que o grau de similaridade mínima (*minSimilarity*) é definido como *0.7*, essa consulta será expandida também para *demanda bioquímica de oxigenio*;
- Proximidade todo-parte: expande a partir do princípio que conceitos constituídos por um conjunto aproximado de partes em comum podem ser semanticamente próximos. Baseados nos relacionamentos da meta-ontologia difusa, *fuz:hasPart* (relação semântica de "todo" para "parte") e *fuz:partOf* (relação semântica de "parte" para "todo"), se estabelece relações de mereologia entre as instâncias da ontologia difusa. Assim, considere como exemplo que a instância *índice de qualidade de água bruta* é expandida para *índice de qualidade de água*, uma vez que possui em comum com esta 9 instâncias de um total de 13 que o constituem, possuindo um grau de proximidade todo-parte (*minCloseness*) maior que o estabelecido pelo usuário no sistema;
- Transitividade: analisa relacionamentos transitivos da ontologia difusa, expandindo a consulta para os termos transitivos ao termo original. Como por exemplo, o relacionamento transitivo entre as bacias, formando uma hierarquia de sub-bacias. Seguindo esse exemplo, considere a *bacia do rio Feijão* uma sub-bacia da *bacia rio Tietê*, e essa por sua vez uma sub-bacia da *bacia do rio Paraná*, então uma consulta à *bacia do rio Paraná* é expandida também para *bacia do rio Tietê* e conseqüentemente para *bacia do rio Feijão*.

Desse modo, alcançou-se como resultado principal do trabalho o desenvolvimento de um Sistema de Integração de Dados com características especiais para integração de dados do domínio de análise de bacias hidrográficas, mas flexível para utilização em outros domínios. Além dessas características, tem como diferencial a utilização

do módulo de expansão da consulta do sistema FOQuE, que permite realizar inferências sobre relações difusas na ontologia.

7.2. Trabalhos futuros

Ainda que desenvolvido um Sistema de Integração de Dados, com as características e funcionalidades básicas definidas e implementadas, este é visto como um protótipo, sendo um primeiro exemplar, um modelo propício a acréscimos e alterações futuras. Essas concepções são deixadas como trabalhos futuros por motivos de prioridades concedidas no desenvolvimento, sendo secundária ao escopo principal do trabalho, ou surgiram com a experiência adquirida no desenvolvimento e testes. Essas propostas, para acréscimos e melhorias ou soluções de problemas em aberto, são listadas a seguir:

- Prover os serviços do BD Geográfico em um *Web Service* [51]. Dessa forma, outros aplicativos, poderão acessar os serviços providos pelo BD Geográfico de forma padronizada, independente das tecnologias usadas nas implementações destes aplicativos;
- Aplicação de técnicas de extração da informação [52]. Ao se trabalhar dentro do domínio de análises de bacias hidrográficas verifica-se que muitas das informações pertinentes a tal domínio estão em linguagem natural (relatórios textuais, tabelas em documentos de texto etc). Assim, acredita-se que o emprego de técnicas de extração da informação traria um aumento substancial na quantidade de dados;
- Mudança de abordagem de integração de virtual para híbrida. Em análise referente a atualizações das fontes de dados do domínio do caso de estudo realizado no IIE, observa-se que muitas das fontes sofrem pouca ou nenhuma alteração, armazenando dados históricos de projetos finalizados. Portanto, a criação de um DW onde esses dados possam ser materializados e consultados diretamente nesta visão materializada, melhoraria o desempenho de consultas, sem perdas na veracidade dos dados, uma vez que as fontes que sofrem alterações e inserções são acessadas por uma visão virtual;
- Melhoria no processamento da consulta. A sugestão anterior já visa um melhor desempenho no processamento da consulta em relação ao seu tempo de resposta. Resta saber se seriam suficientes para resolverem os altos tempos de resposta, que ocorrem em sistemas implantados em máquinas com desempenho de processamento menor; como ocorre no caso de consultas que são realizadas em menos de meio minuto na

máquina de testes descrita na seção 6.4 e em mais de 47 minutos em outra máquina com poder de processamento menor. Outras propostas poderiam surgir com intuito de melhorar o tempo de resposta de consultas, como a definição de mapeamentos menos flexíveis e mais relacionados às estruturas das fontes de dados;

- Tratamento de limpeza de dados [53]. Os casos de problemas no nível de instância relativos à qualidade dos dados, com erros e inconsistências nos dados das bases, não são tratados pelo sistema. Deixa-se para o responsável de inserção dos dados no banco, ou outro agente humano que se relaciona com o processo, a verificação dos dados a serem inseridos no banco. Assim, fica em aberto a aplicação do tratamento de limpeza de dados no momento da consulta, uma vez que o sistema cuida apenas da integração dos dados das fontes, e não da inclusão de dados nestas fontes;
- Solução dos conflitos de precisão e agregação. Esses dois casos de heterogeneidade semântica e estrutural, respectivamente, não têm solução definida no DISFOQuE. Para criar uma forma de tratá-los é necessária a definição de regras que especifiquem a semântica a ser empregada aos dados nos termos causadores desses conflitos.

8. Referências

[1] FERREIRA, A. B. H. **Novo Dicionário Aurélio da Língua Portuguesa**. 3ª ed., revista e atualizada, editora Positivo, 2004.

[2] BRASIL. **Lei n°. 9433, de 8 de janeiro de 1997**. Institui a Política Nacional de Recursos Hídricos, cria o Sistema Nacional de Gerenciamento de Recursos Hídricos, regulamenta o Inciso XIX do ART. 21 da Constituição Federal e altera o ART. 1 da Lei n°. 8001, de 13 de março de 1990, que modificou a Lei n°. 7990, de 28 de dezembro de 1989.

[3] Universidade Estadual de Campinas (UNICAMP). **Instituto de Ecologia da UNICAMP**, 2007 Disponível em: <http://www.eco.unicamp.br/nea/Gestao_Bacia>. Acesso em: 9 set. 2007.

[4] GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: towards a terminological clarification. In: Mars, N (Ed.) **Towards very large knowledge bases**: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995. p. 25-32.

[5] GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge acquisition**. v.5, n. 2, p. 199-220, 1993.

[6] BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. 1997. 243 p. Thesis (PhD). – University of Twente, Enschede.

[7] STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge Engineering: Principles and Methods. **Data & Knowledge Engineering**, v. 25, p. 161-197, 1998.

[8] NOY, N. F.; MCGUINNESS, D. L. **Ontology development 101:** a guide to creating your first ontology. Stanford University, 2001. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

[9] USCHOLD, M.; GRÜNINGER, M. Ontologies and Semantics for Seamless Connectivity. **SIGMOD Record**, v. 33, n. 4, p. 58-64, 2004.

[10] NOVELLO, T C. **Ontologias, sistemas baseados em conhecimento e modelos de banco de dados**. Universidade Federal do Rio Grande do Sul, 2002. Disponível em: <http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/artigo_taisa.pdf>. Acesso em: 12 nov. 2007.

[11] YAGUINUMA, C. A. **Sistema FOQuE para expansão semântica de consultas baseada em ontologias difusas**. Dissertação (Mestrado em Ciência da Computação) Universidade Federal de São Carlos. Maio 2007.

[12] KOIVUNEN, M.; MILLER, E. **W3C semantic web activity**, 2001. Disponível em: <http://www.w3.org/2001/12/semweb-fin/w3csw>. Acesso em: 21 nov. 2007.

[13] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**. p. 28-37, may 2001.

[14] CORCHO, O.; FERNANDEZ, M.; GÓMEZ-PÉREZ, A.; LÓPEZ-CIMA, A**.** Building legal ontologies with METHONTOLOGY and WebODE. **Law and the Semantic Web**. p. 142-157, 2003.

[15] GUARINO, N. Formal ontologies and information systems. In: FIRST INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGIES AND INFORMATION SYSTEMS (FOIS 98), 1998, Trento, Itália. **Proceedings...** Trento: IOS Press, 1998, p. 3-15.

[16] GUARINO, N. Understanding, building and using ontologies. **Journal Human-Computer Studies**. v. 45, n. 2/3, feb./mar. 1997.

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

[18] COSTA, N. C. A. Paraconsistência em informática e inteligência artificial. **Revista dos Estudos Avançados**. v. 14, n. 39, p. 161-174, 2000.

[19] CARNEIRO, M. R. **Ontologias, Web Semântica e aplicações**. Monografia. IME-USP, Instituto de Matemática e Estatística da Universidade de São Paulo. Junho 2003

[20] STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH. **The Protégé ontology editor and knowledge acquisition system**. Disponível em: <http://protege.stanford.edu/>. Acesso em: 14 dez. 2007.

[21] HAKIMPOUR, F.; GEPPERT, A. Resolving semantic heterogeneity in schema integration: an ontology based approach. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS. **Proceedings...** Ogunquit, USA, 2001, p. 297-308.

[22] HALEVY, A. Y. Data integration: a status report. In: CONFERENCE ON DATABASE SYSTEMS FOR BUSINESS TECHNOLOGY AND THE WEB (BTW 2003), 10. **Proceedings...** Germany, 2003.

[23] ÖZSU, M. T.; VALDURIEZ, P. **Principles of distributed database systems**. 2. ed. Englewood Cliffs: Prentice Hall, NJ, 1991.

[24] BUSSE, S. A Specification language for model correspondence assertions, Part I: overlap correspondences. **Forschungsberichte des Fachbereichs Informatik**, Nr. 99-8, TU Berlin. April 1999.

[25] GOH, C. H. **Representing and reasoning about semantic conflicts in heterogeneous information sources**. 1997. Thesis (PhD), Massachusetts Institute of Technology MIT.

[26] SHETH, A.; KASHYAP, V. So Far (schematically) yet so near (semantically). In: IFIP WORKING CONFERENCE ON DATABASE SEMANTICS. **Proceedings...** 1992, p. 283-312.

[27] WIEDERHOLD, G. Mediators in the architecture of future information systems. **IEEE Computer Society Press**. v. 25, p. 38-49, march 1992.

[28] CHAUDHURI, S.; DAYAL, U. An overview of data warehousing and OLAP technology. **SIGMOD Record**, march 1997.

[29] ALASOUD, A.; HAARSLEV, V.; SHIRI, N. A hybrid approach for ontology integration. In: VLDB WORKSHOP ON ONTOLOGIES-BASED TECHNIQUES FOR DATABASES AND INFORMATION SYSTEMS (ODBIS). **Proceedings...**, Trondheim, Norway, september 2005, p. 2-3.

[30] LOSCIO, B. F.; COSTA, T. A.; SALGADO, A. C.; FREITAS, J. S. Query reformulation for an XML-based data integration system. In: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 21., 2006, Dijon - França. **Proceedings...** New York: ACM Press. v. 1. 2006, p. 498-502.

[31] WACHE, H.; VÖGELE, T.; VISSER U.; STUCKENSCHMIDT H.; SCHUSTER G.; NEUMANN H.; HÜBNER S. Ontology-based integration of information - a survey of existing approaches. In: WORKSHOP ON ONTOLOGIES AND INFORMATION SHARING AT THE INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI). **Proceedings...** 2001, p. 108-117.

[32] ARENS Y.; HSU C.; KNOBLOCK C. Query processing in the sims information mediator. **Advanced Planning Technology**. AAAI Press, California, USA, 1996.

[33] BAKER, P. G.; BRASS A.; BECHHOFER S.; GOBLE C.; PATON N.;STEVENS R. Tambis: transparent access to multiple bioinformatics information sources: an overview. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY (ISMB98), 6. **Proceedings...** 1998, p. 25-34.

[34] DECKER, S.; ERDMANN, M.; FENSEL, D.; STUDER, R. Ontobroker: ontology based access to distributed and semi-structured information. In: SEMANTIC ISSUES IN MULTIMEDIA SYSTEMS. **Proceedings...** of DS-8, Rotorua, New Zealand, january 1999, p. 351-369.

[35] PREECE, A.; HUI, K.; GRAY, A.; MARTI, P.; BENCH-CAPON, T.; JONES, D.; CUI, Z. The KRAFT architecture for knowledge fusion and transformation. **Knowledge Based Systems**. v. 13, n. 2-3, p. 113-120, april 2000.

[36] CÂMARA, G.; DAVIS, C.; MONTEIRO, A. M.; D'ALGE, J.C. **Introdução a Ciência da Geoinformação**. 2a. ed., revista e ampliada. São José dos Campos: INPE, 2001. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/introd/>. Acesso em: 08 jan. 2008.

[37] POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostgreSQL – The world´s most advanced open source database**. Disponível em: <http://www.postgresql.org/>. Acesso em: 10 jan 2008.

[38] POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PostGIS**: home. Disponível em: <http://www.postgis.org/>. Acesso em: 10 jan. 2008.

[39] POSTGRESQL GLOBAL DEVELOPMENT GROUP. **PL/pgSQL - SQL Procedural Language**. In: Chapter 37, PostgreSQL 7.4.19 Documentation. Disponível em: <http://www.postgresql.org/docs/7.4/interactive/index.html>. Acesso em: 10 jan. 2008.

[40] AMBLER, S. W. **Choosing a primary key:** natural or surrogate? Disponível em <http://www.agiledata.org/essays/keys.html>. Acesso em: 05 nov 2007.

[41] SUN MICROSYSTEMS, INC. **Java Technology**. Disponível em: <http://java.sun.com/>. Acesso em: 15 jan. 2008.

[42] JENA SEMANTIC WEB FRAMEWORK. Disponível em: <http://jena.sourceforge.net/>. Acesso em: 15 jan. 2008.

[43] SUN MICROSYSTEMS, INC. **Concurrency:** Essential Java Classes, The Java Tutorials. Disponível em: <http://java.sun.com/docs/books/tutorial/index.html>. Acesso em: 15 jan. 2008.

[44] WORLD WIDE WEB CONSORTIUM (W3C). **Extensible Marckup Language (XML)**: W3C Architecture domain. Disponível em: <http://www.w3.org/XML/>. Acesso em: 16 jan. 2008.

[45] DUPAS, F. A.; SOUZA, A. T. S.; TUNDISI, J. G.; TUNDISI, T. M.; ROHM, S. A. Indicadores ambientais para planejamento e gestão do uso de bacias hidrográficas, São Carlos, SP, Brasil. In: **Eutrofização na América do Sul**. São Carlos, 2005, p. 1-1, CD.

[46] U.S. Environmental Protection Agency. **Technical notes on drinking water methods**. Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development. EPA/600/R-94/173. 1994.

[47] Canadian Ministry of the Environment. **Protocol of accepted drinking water testing methods**. Ontario, Canada: Laboratory Services Branch. July 2003.

[48] Companhia de Tecnologia de Saneamento Ambiental – CETESB. **Variáveis de qualidade das águas**. Disponível em: <http://www.cetesb.sp.gov.br/Agua/rios/variaveis.asp>. Acesso em: 30 jan. 2008.

[49] COSTA, T. A. **O gerenciador de consultas de um sistema de integração de dados**. Dissertação (Mestrado em Ciência da Computação). Centro de Informática, UFPE, 2005.

[50] MILLER, R. J.; HERNÁNDEZ M. A.; HAAS, L. M.; YAN, L.; HO, C. T.; FAGIN, R.; POPA, L. The Clio project: managing heterogeneity. **SIGMOD Record**. v. 30, n. 1, p. 78-83, 2001.

[51] BOOTH, D.; HAAS, H.; McCABE, F.; NEWCOMER, E.; CHAMPION, M; FERRIS, C.; ORCHARD, D. **Web Services Architecture**. W3C Working Group Note, 11 february 2004. Disponível em: <http://www.w3.org/TR/ws-arch/>. Acesso em: 07 nov 2007.

[52] SILVA, E. F. A. E.; BARROS, F. A.; PRUDENCIO, R. B. C. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., 2005. ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL DA SBC – ENIA, 5., 2005. **Anais...** São Leopoldo, RS: SBC, v. 1, 2005, p. 504-513.

[53] HERNÁNDEZ, M.; STOLFO, S. Real-world data is dirty: data cleansing and the merge/purge problem. **Data Mining Knowledge Discovery**. v. 2, n. 1, p. 9-37, 1998.

[54] ZIEGLER, P.; DITTRICH, K. R. Three decades of data integration - all problems solved? In: IFIP WORLD COMPUTER CONGRESS (WCC 2004), 18., 2004, Toulouse. **Proceedings…** Toulouse, France, august 2004, p. 3-12.

[55] ALMEIDA, M. B.; BAX, M.P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção . **Ciência da Informação**, Brasilia, v. 32, n. 3, p. 7-20, 2003.

[56] JØRGENSEN, S. E. **Fundamentals of Ecological Modelling**. 2nd ed. Developments in Environmental Modelling, 19. Amsterdam: Elsevier, 1994. 628 p.