



**Alexandre Gazola**

**Uma infra-estrutura de software para alinhamento de  
catálogos heterogêneos**

**Dissertação de Mestrado**

Dissertação apresentada como requisito parcial para  
obtenção do título de Mestre pelo Programa de Pós-  
Graduação em Informática da PUC-Rio.

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro

Março de 2008



**Alexandre Gazola**

## **Uma infra-estrutura de software para alinhamento de catálogos heterogêneos**

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

**Prof. Marco Antonio Casanova**

Orientador  
PUC-Rio

**Prof. Antonio Luz Furtado**

PUC-Rio

**Prof<sup>a</sup>. Karin Koogan Breitman**

PUC-Rio

**Prof. José Eugênio Leal**

Coordenador Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 27 de março de 2008

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

### **Alexandre Gazola**

Graduou-se em Ciência da Computação pela Universidade Federal de Viçosa (UFV) em maio de 2006. Em 2007 foi premiado pelo programa Bolsa Nota 10 da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), destinado aos melhores alunos dos melhores Programas de Pós-Graduação do Estado do Rio de Janeiro. Tem experiência na área de Ciência da Computação, com ênfase em Banco de Dados e Engenharia de Software. É editor técnico da revista Mundo Java, publicação nacional sobre desenvolvimento de software e tecnologia Java.

### Ficha Catalográfica

Gazola, Alexandre

Uma infra-estrutura de software para alinhamento de catálogos heterogêneos / Alexandre Gazola ; orientador: Marco Antonio Casanova. – 2008.

87 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Catálogo de objetos. 3. Esquema. 4. Tesouro. 5. Instância. 6. Alinhamento. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

*A Jesus Cristo, meu melhor amigo*

## **Agradecimentos**

Agradeço primeiramente a Deus, pelo seu imensurável amor derramado na cruz por meio de seu Filho Jesus Cristo. Tudo que sou e tudo que conquistei até hoje provém desse Deus maravilhoso, que sabe dar boas dádivas a seus filhos. Ele é digno de “receber o poder, e riqueza, e sabedoria, e força, e honra, e glória, e louvor”. (Ap 5:12)

Agradeço à minha querida família, pelo conforto e fundamental apoio em todo tempo. Agradeço ao meu pai Nelson, minha mãe Heloísa, meu irmão Rafael e minha irmã Thalita. À minha noiva, Samira, por ter estado sempre ao meu lado, trazendo motivação e carinho, apesar das dificuldades.

Ao meu orientador, o professor Marco Antonio Casanova, pela confiança depositada em mim, por seu auxílio durante este trabalho, pelas oportunidades que me concedeu e por sua paciência.

Aos colegas de curso Algemiro, Fábio Guerra, André Marins e Ricardo Leal. Deixo um agradecimento especial à Daniela Brauner, por ter contribuído de maneira fundamental para que este trabalho se concretizasse.

A todos os colegas do laboratório TecGraf, principalmente ao Vinci Amorim, Roberto Santos, Demétrius Nunes, Eduardo David, Luis Gustavo Ferrão, Luiz Gustavo Ferreira da Silva Costa, Sandro Marinho e Viliam.

À Patrícia, Adriana, Fábio e dona Maria pelo importante suporte que me deram para que eu pudesse me estabelecer no Rio de Janeiro; e à Cecília, por ter me aceitado em sua casa e me proporcionado um ótimo lugar para residir durante meus estudos.

À UFV e ao professor Jugurta Lisboa Filho, pelo aprendizado e pelas oportunidades concedidas.

À CAPES, e à FAPERJ pelo financiamento desta pesquisa.

## Resumo

Gazola, Alexandre; Casanova, Marco Antonio. **Uma infra-estrutura de software para alinhamento de catálogos heterogêneos.** Rio de Janeiro, 2008. 87p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A maior parte dos bancos de dados existentes é projetada de maneira independente e, portanto, é geralmente implementada utilizando diferentes esquemas conceituais, criando um contexto de heterogeneidade em níveis sintático, estrutural e semântico. Não obstante, quando um conjunto de bancos de dados se refere a um mesmo domínio, eventualmente, surge a necessidade de integrá-los em um mesmo banco, ou de intermediar o acesso ao conjunto de bancos de forma transparente. Para tratar o problema da heterogeneidade, torna-se necessário o alinhamento dos esquemas de cada um dos bancos de dados envolvidos. Esse processo geralmente é feito por especialistas de domínio, mas tende a ser um trabalho muito tedioso e propenso a erros. Esta dissertação apresenta o *CatalogMatcher*, uma infra-estrutura de software para alinhamento de catálogos heterogêneos. Um catálogo armazena dados sobre um conjunto de objetos de um determinado domínio, tipicamente classificados por algum tipo de taxonomia ou tesauro. O *CatalogMatcher* contém componentes que implementam estratégias de alinhamento de catálogos heterogêneos utilizando abordagens baseadas em instâncias.

## Palavras-chave

Catálogo de objetos; Esquema; Tesauro; Instância; Alinhamento.

## Abstract

Gazola, Alexandre; Casanova, Marco Antonio. **A software infrastructure for catalog matching**. Rio de Janeiro, 2008. 87p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Most databases are independently designed and, therefore, are usually implemented using different conceptual schemas, which creates a context of syntactic, structural and semantic-level heterogeneity. Nevertheless, when a set of databases refers to a common domain, it may become necessary to integrate them into a single database, or to intermediate access to the databases in a transparent way. To deal with the heterogeneity problem, it becomes necessary to align the conceptual schemas. This process is usually carried out by domain specialists, and tends to be tedious and error-prone. This dissertation presents the *CatalogMatcher*, a software infrastructure for catalog matching. A catalog stores data about a set of objects from a specific domain, typically classified by some sort of taxonomy or thesaurus. The *CatalogMatcher* contains components that implement instance-based alignment strategies.

## Keywords

Objects catalog; Schema; Thesaurus; Instance; Matching.

## Sumário

1 Introdução	13
1.1. Motivação	13
1.2. Trabalhos relacionados	14
1.3. Organização do trabalho	15
2 Fundamentos	17
2.1. Catálogos de objetos	17
2.1.1. Esquemas e instâncias	18
2.1.2. Taxonomias e tesouros	19
2.2. Alinhamento de catálogos	22
2.2.1. Definições básicas	22
2.2.2. Alinhamento baseado em esquema	25
2.2.3. Alinhamento baseado em instâncias	26
2.2.4. Alinhamento baseado em padrões e ontologias	27
2.3. Aplicações de alinhamento de catálogos	29
2.3.1. Sistemas de bancos de dados federados	29
2.3.2. Data warehouses	30
2.3.3. Mediadores	31
3 <i>CatalogMatcher</i> : Uma infra-estrutura de software para alinhamento de catálogos heterogêneos	34
3.1. Estratégias para alinhamento de catálogos	34
3.1.1. Alinhamento de esquemas	34
3.1.2. Alinhamento de tesouros	46
3.2. Arquitetura	49
3.2.1. Módulos principais do <i>CatalogMatcher</i>	49
3.2.2. Módulo básico	52
3.2.3. Módulo de adaptadores de dados para catálogos	54
3.2.4. Módulo de alinhamento de esquemas	56
3.2.5. Módulo de alinhamento de tesouros	60
3.2.6. Módulo de serviços de banco de dados	63



4 Um mediador para catálogos heterogêneos	65
4.1. Casos de uso	65
4.2. Arquitetura	66
4.3. Cenários de uso	69
4.3.1. Mediação de consultas a catálogos de objetos geográficos	70
4.2.1. Mediação de consultas a catálogos de livrarias virtuais	77
5 Conclusões e trabalhos futuros	81
6 REFERÊNCIAS	84

## Lista de figuras

Figura 1 – Taxonomia hierárquica do reino Animália	20
Figura 2 – Fragmento do FTT	21
Figura 3 – Ilustração do alinhamento de dois catálogos fictícios	23
Figura 4 – Uma classificação de abordagens para alinhamento de esquemas (Rahm & Bernstein, 2001)	24
Figura 5 – Grandes grupos de abordagens para alinhamento (Casanova et al., 2007)	24
Figura 6 – Alinhamento sintático de esquemas	26
Figura 7 – Alinhamento semântico de esquemas	26
Figura 8 – Um típico sistema de banco de dados federados	30
Figura 9 – Arquitetura de uma DW	31
Figura 10 – Sistema implementando a arquitetura mediador-adaptador	32
Figura 11 – Técnica de <i>sondagem por consultas</i> utilizada para a construção da matriz de ocorrências	36
Figura 12 – Matriz de ocorrências de $S_1 \times S_2$	37
Figura 13 – Matriz de ocorrências de $S_1 \times S_2$	39
Figura 14 – Alinhamento encontrado	39
Figura 15 – Estado final das classes de alinhamento.	44
Figura 16 – Diagrama de pacotes do <i>CatalogMatcher</i>	51
Figura 17 – Diagrama de classes do módulo básico	54
Figura 18 – Diagrama de classes do módulo de adaptadores de dados para catálogos	56
Figura 19 – Diagrama de classes do módulo de alinhamento de esquemas	59
Figura 20 – Diagrama de classes do módulo de alinhamento de tesouros	61
Figura 21 – Cálculo da taxa de alinhamento entre um termo-fonte e um termo-alvo de dois tesouros	62
Figura 22 – Contagem de instâncias idênticas em dois catálogos de acordo com os termos especificados	63
Figura 23 – Diagrama de classes do módulo de serviços de banco de dados	64
Figura 24 – Diagrama de Casos de Uso	66
Figura 25. Arquivo XML utilizado para o registro de fontes no mediador	66
Figura 26 – Arquitetura do mediador	67
Figura 27 – Esquema conceitual do banco de dados do mediador	69

Figura 28 – Consulta por objetos contendo a palavra-chave “Montanha”	75
Figura 29 – Resposta fornecida pelo mediador considerando o Cenário 1	76
Figura 30 – Resposta fornecida pelo mediador considerando o Cenário 2	76
Figura 31 – Resposta fornecida pelo mediador considerando o Cenário 3	77
Figura 32 – Resposta fornecida pelo mediador considerando o Cenário 1	79
Figura 33 – Resposta fornecida pelo mediador considerando o Cenário 2	80

## Lista de tabelas

Tabela 1 – Fragmento de dados do catálogo da ADL	18
Tabela 2 – Descrição do esquema do catálogo da ADL	19
Tabela 3 – Algumas alternativas para representação de esquemas	19
Tabela 4 – Propriedades de termos de um tesouro	21
Tabela 5 – Conjunto de instâncias de $C_1$ (instâncias globais)	37
Tabela 6 – Conjunto de instâncias de $C_2$	37
Tabela 7 – Catálogos de exemplo para geração do esquema global	43
Tabela 8 – Catálogos de exemplo	48
Tabela 9 – Esquema exportado pela GNS	71
Tabela 10 – Esquema exportado pela GeoNames	72
Tabela 11 – Esquema exportado pela ADL	72
Tabela 12 – Esquema exportado pela Amazon	78
Tabela 13 – Esquema exportado pela Barnes and Noble	78