

**Daniela Francisco Brauner**

## **Alinhamento de esquemas baseado em instâncias**

### **Tese de Doutorado**

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

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro  
Junho de 2008

**Daniela Francisco Brauner**

## **Alinhamento de esquemas baseado em instâncias**

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

**Prof. Marco Antonio Casanova**

Departamento de Informática - PUC-Rio (Orientador)

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

Departamento de Informática - PUC-Rio

**Prof. Ruy Luiz Milidiú**

Departamento de Informática - PUC-Rio

**Prof. Carlos José Pereira de Lucena**

Departamento de Informática – PUC-Rio

**Prof<sup>a</sup>. Cláudia Bauzer Medeiros**

Instituto de Computação - UNICAMP

**Prof<sup>a</sup>. Vânia Maria Ponte Vidal**

Departamento de Computação - UFC

**Prof. José Eugenio Leal**

Coordenador Setorial do Centro  
Técnico Científico - PUC-Rio

Rio de Janeiro, 09 de junho de 2008

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

**Daniela Francisco Brauner**

Mestre em Informática pela PUC-Rio em abril de 2005.  
Graduou-se em Ciência da Computação pela UFPel (Universidade Federal de Pelotas) em março de 2003.

Ficha Catalográfica

Brauner, Daniela Francisco

Alinhamento de esquemas baseado em  
instâncias / Daniela Francisco Brauner ; orientador:  
Marco Antonio Casanova. – 2008.

83 f. : il. ; 30 cm

Tese (Doutorado em Informática)–Pontifícia  
Universidade Católica do Rio de Janeiro, Rio de  
Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Alinhamento de  
esquema. 3. Esquema conceitual. 4. Tesauro. 5.  
Banco de dados. I. Casanova, Marco Antonio. II.  
Pontifícia Universidade Católica do Rio de Janeiro.  
Departamento de Informática. III. Título.

## Agradecimentos

Em primeiro lugar, quero agradecer ao meu orientador, Prof. Casanova, pela confiança que depositou em mim, pelas críticas e sugestões que fez ao longo deste trabalho e, principalmente, pelo apoio constante.

Um agradecimento especial ao Prof. Carlos Lucena pela confiança, orientação e pelas oportunidades concedidas. Agradeço também a alocação no Laboratório de Engenharia de Software, onde encontrei um espaço privilegiado de criatividade, dinamismo e amizade.

Agradeço ao Programa de Pós-Graduação em Informática da PUC-Rio que me permitiu usufruir de sua excelência em ensino e pesquisa, com especial agradecimento ao corpo docente que, por meio de seus cursos, me permitiu aprimorar e ampliar meus conhecimentos na área para desenvolver esta pesquisa.

Agradeço ao CNPq pelo financiamento.

Aos meus amigos, por entenderem a minha ausência e apoiarem as minhas decisões. Em especial à minha grande amiga Vera, por seu exemplo de vida e seus conselhos memoráveis.

Aos meus pais, por me incentivarem a enfrentar os desafios.

Ao meu noivo, Alexandre, por todo carinho, compreensão e principalmente, pela companhia incondicional.

Finalmente, agradeço à Deus!

## Resumo

Brauner, Daniela Francisco; Casanova, Marco Antonio. **Alinhamento de esquemas baseado em instâncias**. Rio de Janeiro, 2008. 83p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Um mediador é um componente de software que auxilia o acesso a fontes de dados. Com o advento da Web, a construção de mediadores impõe desafios importantes, tais como a capacidade de fornecer acesso integrado a fontes de dados independentes e dinâmicas e a habilidade de resolver a heterogeneidade semântica entre os esquemas destas fontes. Para lidar com esses desafios, o alinhamento de esquemas é uma questão fundamental. Nesta tese são propostas abordagens de alinhamento de esquemas de classificação (tesauros) e esquemas conceituais, utilizando instâncias como evidências para os mapeamentos. As abordagens propostas são classificadas em dois tipos: adaptativa e *a priori*, referindo-se, respectivamente, à descoberta dos mapeamentos de forma incremental ou à definição dos mapeamentos antes da implantação do mediador. Por fim, são apresentados experimentos para validação e teste das abordagens propostas.

## Palavras-chave

Alinhamento de esquemas, esquema conceitual, tesauro, banco de dados.

## Abstract

Brauner, Daniela Francisco; Casanova, Marco Antonio. **Instance-based Schema Matching**. Rio de Janeiro, 2008. 83p. DSc. Thesis - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A mediator is a software component that helps accessing data sources. With the advent of the Web, the design of mediators imposes important challenges, such as the ability of providing integrated access to independent and dynamic data sources and the ability of resolving the semantic heterogeneity between different data source schemas. To deal with these challenges, schema matching is a fundamental issue. In this thesis, matching approaches for classification schemas (thesauri) and conceptual schemas are proposed, using instances as evidences for the mappings. The proposed approaches are classified as adaptative and *a priori*, referring to, respectively, the discovery of the mappings in an incremental way or the definition of the mappings before the deployment of the mediator. Finally, experiments to validate and test the proposed approaches are presented.

## Keywords

Schema matching, conceptual schema, thesaurus, database.

## Sumário

1 Introdução .....	11
1.1 Motivação .....	11
1.2 Solução proposta.....	15
1.3 Organização da tese .....	16
2 Conceitos básicos .....	17
2.1 Arquiteturas tradicionais para integração de dados .....	17
2.2 Alinhamentos sintáticos, semânticos e <i>a priori</i> .....	21
3 Alinhamento de tesauros.....	26
3.1 Introdução .....	26
3.2 Abordagem <i>a priori</i> .....	27
3.3 Abordagem adaptativa .....	29
3.4 Validação e testes .....	33
3.5 Considerações .....	45
4 Alinhamento de esquemas conceituais .....	46
4.1 Introdução .....	46
4.2 Abordagem <i>a priori</i> .....	46
4.3 Abordagem adaptativa .....	49
4.4 Validação e testes .....	53
4.5 Considerações .....	65
5 Trabalhos relacionados .....	67
6 Conclusões e trabalhos futuros .....	74
7 Referências bibliográficas .....	81

## Lista de figuras

Figura 1 – Exemplo de heterogeneidade semântica em tesouros.	13
Figura 2 – Exemplo de heterogeneidade semântica em esquemas conceituais.	14
Figura 3 – A operação de alinhamento.	14
Figura 4 – Arquitetura mediador-adaptador.	18
Figura 5 – Elementos básicos de uma data warehouse e o processo ETL.	19
Figura 6 – Alinhamento de esquemas conceituais e transformação de dados.	20
Figura 7 – Exemplos das abordagens sintática e semântica.	22
Figura 8 – Exemplos de instâncias equivalentes.	23
Figura 9 – O processo de alinhamento <i>a priori</i> de tesouros.	27
Figura 10 – O processo de alinhamento <i>adaptativo</i> de tesouros.	29
Figura 11 – Arquitetura proposta para um mediador utilizando a abordagem <i>adaptativa</i> para alinhamento de tesouros.	32
Figura 12 – Processo de validação cruzada e teste.	37
Figura 13 – Gráficos de “Precisão x Limiar” para abordagem de alinhamento <i>a priori</i> ( $T_A \rightarrow T_B$ ).	39
Figura 14 – Gráficos de “Precisão x Limiar” para abordagem de alinhamento <i>a priori</i> ( $T_B \rightarrow T_A$ ).	40
Figura 15 – Gráfico variando o $\alpha$ com “Precisão x Limiar” para abordagem de alinhamento <i>adaptativa</i> ( $T_A \rightarrow T_B$ ).	42
Figura 16 – Gráfico variando o $\alpha$ com “Precisão x Limiar” para abordagem de alinhamento <i>adaptativa</i> ( $T_B \rightarrow T_A$ ).	43
Figura 17 – O processo de alinhamento <i>a priori</i> de esquemas conceituais.	47
Figura 18 – O processo de alinhamento <i>adaptativo</i> de esquemas conceituais.	49
Figura 19 – Arquitetura proposta para um mediador utilizando a abordagem <i>adaptativa</i> para alinhamento de esquemas.	52
Figura 20 – Fragmento do XML de retorno do serviço de consulta do GeoNames.	54
Figura 21 – Fragmento do XML de retorno do serviço de consulta do ADL Gazetteer.	55
Figura 22 – Modelo E-R do esquema global utilizado.	57
Figura 23 – Matrizes de ocorrências ADL Gazetteer X Esquema global.	59
Figura 24 – Matrizes de ocorrências GeoNames X Esquema global.	59
Figura 25 – Matriz de mapeamentos ADL Gazetteer X Esquema global.	60
Figura 26 – Matriz de mapeamentos GeoNames X Esquema global.	60
Figura 27 – Matriz de ocorrências GeoNames X ADL Gazetteer.	64
Figura 28 – Matriz de mapeamentos GeoNames X ADL Gazetteer.	65
Figura 29 – Classificação das técnicas de alinhamento (Euzenat & Shvaiko, 2007).	68



## Lista de tabelas

Tabela 1 – Classes principais dos tesouros GONet e ADL FTT.	34
Tabela 2 – Fragmento do tesouro do GONet.	35
Tabela 3 – Fragmento do tesouro ADL FTT.	35
Tabela 4 – Exemplos de objetos equivalentes extraídos da ADL Gazetteer ( $C_A$ ) e do GONet Names Server ( $C_B$ ).	36
Tabela 5 – Exemplo de um conjunto de validação.	38
Tabela 6 – Exemplo de um conjunto de treinamento.	38
Tabela 7 – Exemplos de objetos equivalentes extraídos da ADL Gazetteer ( $C_A$ ) e do GONet Names Server ( $C_B$ ) e suas classificações segundo $T_A$ e $T_B$ .	40
Tabela 8 – Exemplos dos alinhamentos obtidos via abordagem <i>a priori</i> .	41
Tabela 9 – Exemplos dos alinhamentos obtidos via abordagem <i>adaptativa</i> .	44
Tabela 10 – Esquema de exportação da fonte GeoNames.	56
Tabela 11 – Esquema de exportação da fonte ADL Gazetteer.	56
Tabela 12 – Atributos da classe <i>GeoInstance</i> do esquema global.	58
Tabela 13 – Atributos da classe <i>GeoType</i> do esquema global.	58
Tabela 14 – Fragmento do conjunto de instâncias de referência.	58
Tabela 15 – Co-ocorrência de “Mount Everest” nos resultados do serviço de busca do GeoNames.	60
Tabela 16 – Mapeamentos corretos entre ADL Gazetteer X Esquema global.	61
Tabela 17 – Mapeamentos corretos entre GeoNames X Esquema global.	61
Tabela 18 – Esquema de exportação da fonte ADL Gazetteer.	63
Tabela 19 – Esquema de exportação da fonte GeoNames.	63
Tabela 20 – Mapeamentos corretos entre GeoNames X ADL Gazetteer.	64
Tabela 21 – Tabela comparativa dos trabalhos relacionados.	73

## Lista de siglas

ADL	Alexandria Digital Library
ADL FTT	Alexandria Digital Library Feature Type Thesaurus
DW	Data Warehouse
ER	Entidade-Relacionamento
ESRB	Entertainment Software Rating Board
ETL	Extract - Transformation - Load
GEOnet	GEOnet Names Server
GeoNames	GeoNames geographical database
GTIN	Global Trade Item Number
HTML	HyperText Markup Language
ISBN	International Standard Book Number
ISO	International Standards Organization
MAFL	Módulo de Acesso às Fontes Locais
MAFR	Módulo de Acesso às Fontes Remotas
MEM	Módulo de Esquema Mediado
METM	Módulo Estimador de Taxas de Mapeamento
MGC	Módulo Gerente de Consultas
MI	Módulo de Interface
MR	Módulo de Registro
NGA	National Geospatial Intelligence Agency
OWL	Ontology Web Language
RDF	Resource Description Framework
SGBD	Sistema Gerenciador de Bancos de Dados
SGM	Sistema de Gerenciamento de Modelos
SQL	Structured Query Language
TI	Tecnologia da Informação
UPC	Universal Product Code
W3C	World Wide Web Consortium
XML	eXtensible Markup Language