



**André Luiz Almeida Marins**

## **Modelos Conceituais para Proveniência**

### **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



**André Luiz Almeida Marins**

## **Modelos Conceituais para Proveniência**

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

Departamento de Informática – PUC-Rio

**Prof. Antonio Luz Furtado**

Departamento de Informática – PUC-Rio

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

Departamento de Informática – PUC-Rio

**Prof<sup>a</sup>. Melissa Lemos Cavalière**

Departamento de Informática – PUC-Rio

**Prof. José Eugenio Leal**

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de Janeiro, 18 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.

## **André Luiz Almeida Marins**

Graduado em Engenharia de Computação pela Pontifícia Universidade Católica do Rio de Janeiro (1996), com especialização em Gestão Executiva no Programa de Desenvolvimento Gerencial (PDG Senior Exec) pelo Instituto Brasileiro de Mercado de Capitais (2001), Atualmente é pesquisador do Laboratório Tecgraf e atua nas áreas de Banco de Dados, Proveniência e Rastreabilidade.

### Ficha Catalográfica

Marins, André Luiz Almeida

Modelos Conceituais para Proveniência / André Luiz Almeida Marins; orientador: Marco Antonio Casanova. – 2008.

189 f.; 30 cm

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

Inclui bibliografia.

1. Informática – Teses. 2. Proveniência. 3. Rastreabilidade. 4. Modelagem Conceitual. 5. Alinhamento de Ontologias. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

## Agradecimentos

Aos patrocinadores CNPq e PUC-Rio pelas bolsas de fomento e isenção, obrigado. À Noemi, Hugo, Paulo, Costa, Gustavo, Fábio, Rosana e Eduardo pelas referências creditadas e a eterna amizade, muito obrigado!

Um provérbio árabe que adaptei diz que há quatro tipos de pessoas: aquelas que não sabem e não identificam o que não sabem (tolas); algumas que não sabem e reconhecem que não sabem (humildes); outras que sabem, mas não compreendem o que sabem (aprendizes); e as que sabem e conhecem o que sabem (mestres).

Às vezes, precisei me perder para me encontrar. Percebi que não sabia e, doutores-filósofos me indicaram a direção. Lutei, persisti e corrigi o percurso. Ao Marco (Casa) - professor, orientador e amigo - agradeço os ensinamentos práticos e teóricos, pessoais e profissionais.

(Re)conheci pessoas incríveis que levarei comigo por toda a vida. Elegi duas que representam com louvor todas as demais não mencionadas. Daniela e Ricardo, vocês são demais.

Nas horas mais difíceis assegure. Nas fáceis ratifique. Em ambas, reifique: o amor, o afeto, a amizade. O produto vira serviço exclusivo, terra fértil e água fresca para felicidade que encanta produtor e consumidor. Também registre, registre muito, registre tudo. Para que na falta de memória, a história seja preservada. Para aprender e ensinar depois de apreender. Para sorrir ou chorar, mas viver. Obrigado a minha família, minha base. Renata, você é única, voemos juntos. Rosely e Hélio, minha genética e educação tem raízes nobres em vocês. Rachel e Zuleika, duas gerações diferentes que complementam a minha e, com elas sou muito, mas muito, muito mais feliz.

A todos que contribuíram para o sucesso desta pesquisa reitero: tenho a certeza que tudo é efêmero. Não obstante, as experiências que vivi, estão em meu coração e as levarei sempre comigo. Obrigado por tudo! Vamos em frente!

## Resumo

Marins, André Luiz Almeida; Casanova, Marco Antonio. **Modelos Conceituais para Proveniência**. Rio de Janeiro, 2008. 189p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sistemas de informação, desenvolvidos para diversos setores econômicos, necessitam com maior frequência de capacidade de rastreabilidade dos dados. Para habilitar tal capacidade, é necessário modelar a proveniência dos dados. A proveniência permite testar conformidade com a legislação, repetição de experimentos e controle de qualidade, entre outros. Habilita também a identificação de participantes (determinantes ou aderentes) como pessoas, organizações, agentes de software entre outros e, permite associá-los a atividades, eventos ou processos. Pode ser utilizada para estabelecer níveis de confiança para as transformações dos dados. Esta dissertação propõe um modelo genérico de proveniência criado com base no alinhamento de recortes de ontologias de alto nível, padrões internacionais e propostas de padrões que tratam direta ou indiretamente de conceitos relacionados à proveniência. As contribuições da dissertação são, portanto em duas direções: um modelo conceitual para proveniência - bem fundamentado - que facilita a interoperabilidade e a aplicação da estratégia de projeto conceitual baseada em alinhamento de ontologias.

## Palavras-chave

Proveniência; rastreabilidade; modelagem conceitual; alinhamento de ontologias.

## Abstract

Marins, André Luiz Almeida; Casanova, Marco Antonio. **Provenance Conceptual Models**. Rio de Janeiro, 2008. 189p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Information systems, developed for several economic segments, increasingly demand data traceability functionality. To endow information systems with such capacity, we depend on data provenance modeling. Provenance enables legal compliance, experiment validation, and quality control, among others. Provenance also helps identifying participants (determinants or immanents) like people, organizations, software agents among others, as well as their association with activities, events or processes. It can also be used to establish levels of trust for data transformations. This dissertation proposes a generic conceptual model for provenance, designed by aligning fragments of upper ontologies, international standards and broadly recognized projects. The contributions are in two directions: a provenance conceptual model - extensively documented – that facilitates interoperability and the application of a design methodology based on ontology alignment.

## Keywords

Provenance; traceability; conceptual modeling; ontology alignment.

## Sumário

1	Introdução	15
1.1.	Motivação	16
1.2.	Problema	18
1.3.	Metodologia	20
1.4.	Objetivo	22
1.5.	Organização do Texto	23
2	Contexto para Proveniência	25
2.1.	O Conceito de Proveniência	25
2.1.1.	Definição de Dicionário	26
2.1.2.	Proveniência como Metadado	28
2.1.2.1.	Dublin Core	29
2.1.2.2.	Warwick Framework	29
2.1.2.3.	ISO 19115:2003	30
2.1.2.4.	Linhagem	30
2.1.2.5.	Anotação	30
2.1.3.	Mapeamento de Conceitos de História	31
2.1.4.	Proveniência e Ciclo de Vida	34
2.1.5.	A Sétima Pergunta de Proveniência	35
2.1.6.	Ontologia Parcial para Proveniência (prévia 1)	36
2.2.	Aspectos de Projetos baseados em Proveniência	37
2.3.	Ontologias de Alto Nível	41
2.3.1.	Preliminares	41
2.3.2.	DOLCE	42
2.3.3.	SUMO	44
2.3.4.	OPENCYC	46
2.3.5.	COSMO e o Alinhamento entre Ontologias de Alto Nível	48
2.3.6.	Ontologia Parcial para Proveniência (prévia 2)	49
2.3.7.	Considerações Finais	50
2.4.	Padrões e Projetos cobrindo o conceito de proveniência	51
2.4.1.	Preliminares	51
2.4.2.	Padrões	52
2.4.2.1.	ISO 14721:2003 (OAIS)	52

2.4.2.2. ISO 21127:2006 (CIDOC CRM)	57
2.4.2.3. Alinhamento entre Padrões e Ontologias de Alto Nível	68
2.4.3. Projetos	69
2.4.3.1. FRBRoo	70
2.4.3.2. INDECS	74
2.4.3.3. Harmony	79
2.4.3.4. Alinhamento entre Projetos e o Padrão ISO 21127	83
2.4.3.4.1. FRBRoo / CIDOC CRM	84
2.4.3.4.2. Harmony / CIDOC CRM	89
2.4.4. Ontologia Parcial para Proveniência (prévia 3)	93
2.4.5. Considerações Finais	94
3 Modelo de Proveniência	95
3.1. Preliminares	95
3.2. Estratégia de Projeto (Proveniência no Centro)	96
3.3. Tática de Projeto (PPCO)	97
3.4. Metaproveniência	99
3.5. Modelo Conceitual	110
3.5.1. Modelo Mínimo (3P)	111
3.5.1.1. Classes	115
3.5.1.2. Propriedades	122
3.5.2. Expansões	123
3.5.2.1. PO ( <i>Partial Order</i> )	123
3.5.2.2. TR ( <i>TRaceability – TRace and TRack</i> )	126
3.5.2.2.1. RQR ( <i>Reification of Qualified Relations</i> )	126
3.5.2.2.2. DISP ( <i>Determinant, Immanent, Source Product</i> )	128
3.5.2.3. CH ( <i>Cultural Heritage</i> )	130
3.6. Considerações Finais	135
4 Aplicações do Modelo de Proveniência	138
4.1. Preliminares	138
4.2. Aplicações de Desktop Semântico	139
4.3. Generalização para Design Rationale	144
4.4. Centro de Informações	147
4.5. Serviço de Proveniência na Web para ferramenta de Gerenciamento de Configuração de Software	151
4.5.1. Especificação do Serviço	152



4.5.1.1. Requisitos	152
4.5.1.2. Descrição das Operações	153
4.5.1.3. Arquitetura	154
4.5.2. Ambiente de Desenvolvimento	155
4.5.2.1. Catálogo TDK	155
4.5.2.2. Modelo Físico construído por Reuso e Adaptação	156
4.5.2.3. Abstração para o Modelo Físico - Ferramenta BI	159
4.5.3. Estudo da Ferramenta Trac	160
4.5.3.1. Preliminares	160
4.5.3.2. Mapeamento para o Modelo de Proveniência	165
4.5.3.3 Respostas do Serviço de Proveniência na Web	167
4.6. Considerações Finais	176
5 Conclusões	177
6 Referências Bibliográficas	182

## Lista de figuras

Figura 1: Rastreabilidade na Cadeia Produtiva baseado em (Bechini et al., 2005).....	16
Figura 2: Ontologia parcial para proveniência .....	21
Figura 3: Visão de História de Mario Bunge (Ram, 2006a) .....	32
Figura 4: Ciclo de vida da informação (Ram, 2006a) .....	34
Figura 5: Ontologia parcial para proveniência (prévia 1).....	36
Figura 6: Taxonomia de Aspectos de Proveniência de Dados (Simmhan et al., 2005).....	38
Figura 7: Hierarquia de especialização parcial da ontologia DOLCE (Semy et al., 2004).....	43
Figura 8: Hierarquia de especialização parcial da ontologia SUMO (Semy et al., 2004).....	45
Figura 9: Hierarquia de especialização parcial da ontologia OpenCyc (Semy et al., 2004).....	47
Figura 10: Ontologia parcial para Proveniência (prévia 2) .....	49
Figura 11: Detalhamento de informação de contexto (Lavoie et al., 2002).....	55
Figura 12: Detalhamento de informação de proveniência (Lavoie et al., 2002).....	55
Figura 13: Algumas classes do CIDOC CRM e destaque para o evento no centro .....	58
Figura 14: Taxonomia para E2 Temporal Entity (Doerr, 2005).....	60
Figura 15: Algumas classes da ISO 21127:2006 expressas com elementos do Dublin Core baseado em (Doerr, 2005).....	62
Figura 16: Representação Gráfica do CIDOC CRM Core DTD (Sinclair et al., 2006a).....	65
Figura 17: Exemplo utilizando o CRM Core (Sinclair et al., 2006a) .....	66
Figura 18: Exemplo XML do CIDOC CRM Core (Sinclair et al., 2006a).....	68
Figura 19: Relação entre projetos estudados e a ISO 21127:2006.....	70
Figura 20: Visão Geral do Modelo FRBR baseado em (FRBR Review Group, 1998).....	71
Figura 21: Instanciações para o Grupo 1 baseado em (Tillett, 2003).....	72
Figura 22: Principais classes e relacionamentos INDECS adaptado de (Bearman et al., 1999) .....	75

Figura 23: Visão centralizada da classe Evento no modelo INDECS adaptada de (Rust & Bide, 2000) .....	78
Figura 24: Taxonomia para Modelo ABC baseado em (Doerr et al., 2003) .....	80
Figura 25: Exemplo de Instanciação do modelo ABC (Lagoze & Hunter, 2001).....	82
Figura 26: Alinhamento de classes dos modelos FRBRoo e CIDOC CRM (ISO 21127:2006) baseado em (Bouef & Doerr, 2007) .....	85
Figura 27: Alinhamento parcial das classes dos modelos ABC e CIDOC CRM (ISO 21127:2006) baseado em (Doerr et al., 2003) .....	90
Figura 28: Alinhamento parcial de propriedades entre modelos ABC e CIDOC CRM (ISO 21127:2006) baseado em (Doerr et al., 2003).....	92
Figura 29: Ontologia parcial para Proveniência (prévia 3) .....	93
Figura 30: CODeP Participation reificado (Gangemi, 2005).....	99
Figura 31: Padrão de Ontologia D&S proposto em (Masolo et al., 2003) .....	101
Figura 32: CODeP DnS simplificado em (Gangemi, 2006) .....	102
Figura 33: Entidades do modelo INDECS (Rust & Bide, 2000).....	106
Figura 34: A classe E52 Time-Span, outras classes e relacionamentos da ISO 21127:2006.....	112
Figura 35: Diagrama UML com principais classes e relacionamentos do modelo conceitual e ilustrado com conceitos abstratos ( <i>Wh-questions</i> ) ...	114
Figura 36: Rastreabilidade na Cadeia Produtiva baseado em (Bechini et al., 2005).....	124
Figura 37: Três notações gráficas para especificação de procedimentos (Sowa, 2001a) .....	126
Figura 38: Arquitetura Simplificada para Desktop Semântico .....	141
Figura 39: Arquitetura de Busca em Desktop Semântico baseada no Beagle <sup>++</sup> .....	142
Figura 40: Resultado de uma consulta semântica.....	143
Figura 41: Conceitos de Proveniência ajudam a organizar e classificar o resultado de uma consulta semântica. ....	144
Figura 42: Modelo Kuaba e conceitos abstratos adaptado de (Medeiros, 2006).....	145
Figura 43: Modelo Funcional de um Centro de Informações adaptado da ISO 14721:2003.....	148
Figura 44: Arquitetura do <i>Web Provenance Service</i> (WPS).....	155
Figura 45: Principais Tabelas do Catálogo TDK Adaptado .....	158

Figura 46: Abstração BI para o modelo físico.....	160
Figura 47: Tela de Entrada do Trac.....	161
Figura 48: Abstração BI para o esquema do banco de dados do Trac .....	162
Figura 49: Rastreabilidade de <i>ChangeSets</i> .....	163
Figura 50: Referências cruzadas entre <i>Ticket</i> e <i>ChangeSet</i> apresentada no <i>Ticket</i> .....	164
Figura 51: Referências cruzadas entre <i>ChangeSet</i> e <i>Ticket</i> apresentada no <i>Timeline</i> .....	164
Figura 52: Página de <i>Roadmap</i> da ferramenta Trac que apresenta os <i>milestones</i> existentes no projeto.....	168
Figura 53: Resultado com todos os eventos no banco de dados de proveniência. ....	170
Figura 54: Eventos de <i>Commit</i> do Subversion que correspondem a <i>ChangeSets</i> .....	172
Figura 55: <i>ChangeSet</i> 1129 que concluiu o <i>ticket</i> #20.....	173
Figura 56: <i>Tickets</i> do <i>Milestone</i> Branch 1.5 .....	174
Figura 57: <i>Tickets</i> #4, #9 e #20 e respectivos IDs de participantes .....	175

## Lista de tabelas

Tabela 1: Mapeamento entre conceitos de História e Proveniência (Ram, 2006a)	31
Tabela 2: Alinhamento parcial de Ontologias de Alto Nível	48
Tabela 3: Exemplo de um pacote de informação (informação de conteúdo + informação de descrição de preservação) (NBR 15472:2007)	54
Tabela 4: Propriedades que apresentam as classes Evento, Agente e Ação como classe-domínio ou classe-imagem (ISO 21127 2006).	61
Tabela 5: Alinhamento parcial sugerido entre ontologias de alto nível e padrões ISO adaptado de (Casanova et al., 2007)	69
Tabela 6: Alinhamento parcial entre modelo INDECS e FRBR baseado em (Rust, 2005)	76
Tabela 7: Detalhamento da interpretação de abstrações e manifestações do modelo INDECS baseado em (Rust, 2005)	77
Tabela 8: Conceitos de proveniência do modelo INDECS	78
Tabela 9: Alinhamento parcial entre os modelos ABC e Dublin Core adaptado de (Hunter, 2000)	82
Tabela 10: Alinhamento parcial de classes entre modelos FRBRoo e CIDOC CRM (Doerr et al., 2007)	85
Tabela 11: Hierarquia de especialização do modelo FRBRoo para a classe F11 Event	86
Tabela 12: Propriedades que representam a importância da classe Evento no modelo FRBRoo	87
Tabela 13: Propriedades que representam a importância da classe Evento no modelo FRBRoo (continuação da Tabela 12)	88
Tabela 14: Alinhamento parcial de classes entre modelos ABC e CIDOC CRM (Doerr et al., 2003)	90
Tabela 15: Alinhamento parcial de propriedades entre modelos ABC e CIDOC CRM (Doerr et al., 2003)	91
Tabela 16: Propriedades que capturariam parcialmente a noção de razão ( <i>Reason</i> )	106
Tabela 17: Mapeamento de conceitos abstratos de proveniência	108
Tabela 18: Hierarquia parcial de especialização após descarte dos conceitos abstratos	109

Tabela 19: Descrição das Propriedades do Modelo de Proveniência	122
Tabela 20: Descrição das Propriedades do Modelo de Proveniência (parte 1 de 2)	134
Tabela 21: Descrição das propriedades da expansão (parte 2 de 2)	135
Tabela 22: Analogia entre Design Rationale e Proveniência	146
Tabela 23: Mapeamento entre conceitos do Trac em classes do modelo de proveniência	166
Tabela 24: Consulta SQL ( <i>prepared query</i> ) para a operação <i>getObject</i> que especifica o tipo do objeto que deve ser retornado	171
Tabela 25: Linguagem para alinhamento entre ontologias (Scharffe & Bruijn, 2005)	179