

Mesodata:  
Engineering Domains for Attribute  
Evolution and Data Integration

by

Denise Bernadette Angela de Vries, *B.Comp.& Inf.Sc., B.Sc.(Hons)*  
School of Informatics and Engineering,  
Faculty of Science and Engineering

21 October 2005

A thesis presented to the  
Flinders University of South Australia  
in total fulfillment of the requirements for the degree of  
Doctor of Philosophy

# Contents

<b>Abstract</b>	<b>ix</b>
<b>Certification</b>	<b>xi</b>
<b>Acknowledgements</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>5</b>
2.1 Database Organisation and Evolution . . . . .	5
2.2 Causes of Change . . . . .	6
2.3 Change Management . . . . .	7
2.4 Information Capacity . . . . .	9
2.5 Techniques for Database Evolution . . . . .	12
2.6 Schema Integration . . . . .	12
2.6.1 Common Data Model . . . . .	13
2.6.2 Schema Intension Graphs . . . . .	14
2.6.3 Hypergraph Data Model . . . . .	15
2.6.4 Evolutionary ER Diagrams . . . . .	17
2.6.5 Schematic Conflicts . . . . .	17
2.7 Schema Matching . . . . .	19
2.8 Semantic Heterogeneity . . . . .	20
2.9 Object-Relational Databases . . . . .	21
2.10 Data Conversion . . . . .	23
2.10.1 Attribute Evolution . . . . .	23
2.11 Data and View Integration . . . . .	25

2.12	Mediation Techniques . . . . .	26
2.13	Ontologies . . . . .	31
2.14	Concept Graphs . . . . .	33
2.15	Knowledge Interchange . . . . .	34
2.16	Summary . . . . .	35
<b>3</b>	<b>Mesodata in DBMS</b>	<b>38</b>
3.1	Modelling . . . . .	38
3.2	Mesodata . . . . .	41
3.3	Mesodata Domains . . . . .	42
3.3.1	Definition of the Mesodata Domain . . . . .	43
3.3.2	Extended Querying . . . . .	44
3.4	Structured Domains . . . . .	45
3.4.1	Filters . . . . .	46
3.4.2	Topological Spaces . . . . .	46
3.5	Mesodata Operators . . . . .	49
3.6	Comparison of Mesodata with User-Defined Types . . . . .	49
3.7	Conceptual Model Incorporating Mesodata . . . . .	51
3.8	Summary . . . . .	53
<b>4</b>	<b>Reference Data Language</b>	<b>55</b>
4.1	Aims . . . . .	55
4.2	Mesodata Definition Language . . . . .	56
4.2.1	Create Domain Syntax . . . . .	56
4.2.2	Drop Domain Syntax . . . . .	57
4.2.3	Alter Domain Syntax . . . . .	57
4.2.4	Refresh Domain Syntax . . . . .	58
4.2.5	Describe Domain Syntax . . . . .	58
4.2.6	Show Domains Syntax . . . . .	58
4.3	Mesodata Extended SQL . . . . .	59
4.3.1	Create Table Syntax . . . . .	59
4.3.2	Alter Table Syntax . . . . .	59
4.3.3	Drop Table Syntax . . . . .	60

<i>CONTENTS</i>	iii
4.3.4 Describe Mesodata Type Syntax . . . . .	60
4.3.5 Show Mesodata Types Syntax . . . . .	61
4.4 Extensions to Manipulation Language . . . . .	61
4.4.1 Select Syntax . . . . .	61
4.5 Summary . . . . .	63
<b>5 Application of Mesodata</b>	<b>64</b>
5.1 Domain Evolution . . . . .	64
5.2 Change Management . . . . .	65
5.3 Attribute Domain Evolution . . . . .	66
5.4 Categories of Domain Evolution . . . . .	68
5.4.1 Attribute Representation Change . . . . .	69
5.4.2 Domain Constraints Change . . . . .	69
5.4.3 Domain Perception (meaning) Change . . . . .	70
5.4.4 Minimise Change . . . . .	70
5.5 Data Integration . . . . .	71
5.6 Enhanced Queries . . . . .	73
5.6.1 Example of a Circular Domain . . . . .	74
5.7 An Object-Relational Example . . . . .	75
5.7.1 Hierarchical Domain . . . . .	77
5.8 Summary . . . . .	77
<b>6 Empirical Study of a Database System</b>	<b>79</b>
6.1 Motivation for the Study . . . . .	79
6.2 System Overview and Evolution . . . . .	80
6.2.1 System Metrics . . . . .	81
6.2.2 Stable Characteristics . . . . .	85
6.2.3 Deleted Values . . . . .	85
6.2.4 Modified Domains . . . . .	86
6.3 Data Conversion and Maintenance . . . . .	88
6.4 Summary . . . . .	91

<b>7</b>	<b>Prototype Model</b>	<b>93</b>
7.1	Prototype Evaluation . . . . .	93
7.1.1	Evaluation Criteria . . . . .	94
7.1.2	Prototype Platform . . . . .	94
7.1.3	Prototype Components . . . . .	95
7.1.4	Query Parser . . . . .	96
7.2	Example Database . . . . .	97
7.2.1	Evaluation of Model . . . . .	98
7.2.2	Enhanced Querying . . . . .	99
7.2.3	Domain Perception Change . . . . .	101
7.2.4	Domain Constraints Change . . . . .	103
7.2.5	Data Integration . . . . .	104
7.2.6	Attribute Representation Change . . . . .	104
7.3	Summary of Evaluation . . . . .	105
<b>8</b>	<b>Conclusions and Further Research</b>	<b>106</b>
8.1	Database Evolution . . . . .	106
8.2	Techniques for Database Evolution . . . . .	107
8.3	Data Integration . . . . .	107
8.4	Mesodata Layer . . . . .	108
8.5	Future Research . . . . .	108
8.5.1	DB Platform Support for Mesodata . . . . .	108
8.5.2	XML . . . . .	109
8.5.3	Ontologies of Data Structures . . . . .	109
8.5.4	Mesodata types based on UDTs . . . . .	109
8.5.5	Modelling Tools . . . . .	109
8.5.6	Other Database Technologies . . . . .	109
	<b>Appendices</b>	<b>110</b>
<b>A</b>	<b>Publications Resulting From This Thesis</b>	<b>110</b>
<b>B</b>	<b>Sample Session</b>	<b>114</b>

<b>C</b>	<b>Sample Session SQL Files</b>	<b>139</b>
C.1	adjColours.sql . . . . .	139
C.2	categories.sql . . . . .	140
C.3	shadescolours.sql . . . . .	140
C.4	furnitureB.sql . . . . .	141
C.5	customers.sql . . . . .	142
C.6	suppliers.sql . . . . .	143
C.7	sales.sql . . . . .	144
C.8	salesitem.sql . . . . .	144
C.9	hexColours.sql . . . . .	145
C.10	furnitureC.sql . . . . .	146
C.11	salesB.sql . . . . .	146
C.12	salesitemB.sql . . . . .	147
C.13	codelist.sql . . . . .	147
<b>D</b>	<b>Prototype Functionality</b>	<b>149</b>
<b>E</b>	<b>Prototype Domain Querying</b>	<b>152</b>
<b>F</b>	<b>Data type Comparisons</b>	<b>160</b>
<b>G</b>	<b>Mapping MySQL to Java types</b>	<b>165</b>
	<b>Bibliography</b>	<b>167</b>

# List of Figures

2.1	Techniques for Database Information Systems . . . . .	12
2.2	SIG - Schema Intension Graph . . . . .	15
2.3	HDM - Two Source Schemas and One Global Schema . . . . .	16
2.4	Database Evolution and Related Research Areas . . . . .	35
3.1	A Matrix for Classifying DBMS . . . . .	40
3.2	Mesodata Layer Between Metadata and Data . . . . .	41
3.3	Hierarchy of Some Suggested Mesodata Types for Different Do- main Structures . . . . .	42
3.4	A Filter in Metric Space . . . . .	47
3.5	ERD An Attribute Referencing a Mesodata Domain . . . . .	52
3.6	ERD Multiple Attributes Referencing Mesodata Domains . . . . .	52
3.7	ERD An Attribute Referencing Joined Mesodata Domains . . . . .	52
3.8	ERD An Attribute Referencing Multiple Mesodata Domains . . . . .	53
3.9	ERD UML Notation . . . . .	53
5.1	Ranges of Year Domains . . . . .	67
5.2	Heterogenous but Similar Schemata . . . . .	71
5.3	Example Colour Chart as Weighted Graph . . . . .	72
5.4	Attribute ‘Colour’ Referencing Mesodata Type Weighted Graph . . . . .	73
5.5	Days of the Week with English and French Terms . . . . .	75
5.6	Configurations for Telephone Numbers . . . . .	76
6.1	Schematic of the Database System . . . . .	81
6.2	Growth of Relations . . . . .	82
6.3	Growth of Attributes . . . . .	83
6.4	New Relations . . . . .	83

6.5	Attribute Movement . . . . .	84
6.6	Unmodified Attributes . . . . .	86
6.7	Deleted Attributes . . . . .	87
6.8	Modified Attributes . . . . .	87
6.9	Modified Relations . . . . .	88
7.1	Deployment Diagram . . . . .	96
7.2	Activity Diagram for Mesodata Wrapper . . . . .	97
7.3	Entity-Relationship Model of Test Database . . . . .	98



# List of Tables

2.1	Schematic Conflicts . . . . .	18
2.2	Evolutionary Operations in ORDBs . . . . .	22
2.3	Schematic Changes that Affect Data . . . . .	36
3.1	Partial List of Mesodata Types with Extended SQL Operators . . . . .	48
3.2	Comparison of User Defined Types (UDT) and Mesodata Types . . . . .	50
6.1	Analysis Results . . . . .	92
D.1	Prototype Functionality . . . . .	149
E.1	Prototype Querying . . . . .	153
F.1	Comparison of Data Types . . . . .	161
G.1	Mapping SQL and Java data types (MySQL 2003) . . . . .	165

# Abstract

The introduction of databases for data storage and handling revolutionised the way we dealt with records and enabled simple and fast information processing, aggregation and summarisation. Database and information technology systems have evolved from simple file processing systems to powerful database systems. Data management technology has progressed from hierarchical and network systems to relational databases, data modelling tools and indexing and organisational techniques. The development of Relational Database Management Systems and automated systems put the layout and *form* into the unchanging metadata and gave us *record once* systems.

Unfortunately, the ‘real world’ upon which databases are modelled constantly changes. These changes may affect the schema for a variety of reasons including;

- Unanticipated requirements,
- A change in the universe of discourse,
- A change to the interpretation of facts about the universe of discourse,
- Changes in the form of updates to effect upgrades to the functionality or scope of a system,
- Changes in the form of updates to effect efficiency improvements,
- Changes caused by system operation,
- Error correction.

Different formalisms have been developed to deal with schema changes with the aim being to preserve information capacity and preserve semantic correctness. Schematic changes may be the result of evolving one system or may arise due to the need for merging two or more systems. Schematic conflicts occur which must be resolved and the schemata unified to produce a new version. To reach this goal there are graph based *schema integration* architectures, as well as, semi-automatic systems applying *schema matching* and *schema translation* techniques. These systems also utilise ontologies, thesauri, and so forth to integrate data from heterogeneous sources in order to process queries and views.

Data integration or conversion remains a partially resolved issue. Some metadata changes are managed by changes to application code and system down time for conversion procedures. However an attribute change may result in data loss, changed accuracy, and altered semantics. Whilst the use of ontologies, concept graphs and other knowledge interchange techniques are alleviating the problems of data integration, these structures are not yet an integral part of the database architecture.

This thesis argues a three-level architecture for relational databases with an interface positioned between data and metadata for complex domains. This intermediary level is the *mesodata* layer. This mesodata layer, separate from the metadata and data, provides complex structures, such as graphs, queues, and circular lists, in which to store domain values and their inter-relationships as well as supplying the ‘intelligence’ required to operate and manipulate them. The domain structures enable different orderings that form the bases of filters for enhanced querying and information retrieval. DBMS supplied mesodata types would allow for the re-usable inclusion of domain information such as in ontologies, taxonomies and concept graphs that to date have been only application specific.

# Certification

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

As requested under Clause 14 of Appendix E of the *Flinders University Research Higher Degree Policies and Procedures Manual* I hereby agree to waive the conditions referred to in Clause 13(b) and (c), and thus

- Flinders University may lend this thesis to other institutions or individuals for the purpose of scholarly research;
- Flinders University may reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signed

Dated

Denise Bernadette Angela de Vries

# Acknowledgements

I would like to thank my supervisor Professor John Roddick for his advice, support and enthusiasm throughout my candidature.

There are many people in the School of Informatics and Engineering at Flinders University who have helped me in large ways and small, I believe I owe each person thanks. In particular, the members of the Knowledge Discovery and Intelligent Systems Group for constructive criticism, rigorous discussions and friendship, – (in room number order) Darin Chan, Dongqiang Yang, Trent Lewis, Martin Luerssen, Richard Leibbrandt, Darius Pfitzner, David Powers, Aaron Ceglar, Carl Mooney, Sally Rice, Anna Shillabeer, Edi Winarko, Ron Porter, Paul Calder, Amos Omondi, Tiffany Winn, and Lorraine Harker – and Murk Bottema and Jalina Widjaja for their assistance and comments.

I appreciate too the Flinders Postgraduate Students' Association for providing support, advice, resources and the research training courses and workshops that were so helpful at the beginning of my candidature. Thank you Leonie Randall and Audrey Nicholson.

I am very grateful for the all the assistance I received from Versatile Solutions Pty. Ltd, especially to Mr Arthur Verster for his time and effort.

However, none of this work could have been achieved without the unstinting support of Bart de Vries who must be the most generous, patient and caring person in the world.

Denise Bernadette Angela de Vries  
October 2005  
Adelaide.