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

Adelaide, South Australia, 2006

Contents

List of Figures

List of Tables

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.