INTECoM:

An integrated conceptual data modelling framework.

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Information Systems

at Massey University, Palmerston North

New Zealand

Clare Frances Atkins

2000

Errata Sheet

Page	Location	Action	Content
6	Para 3 - line 2	Insert Footnote after word NIAM	Originally standing for Nijssen's Information Analysis Method, it has been renamed at Nijssen's own request and this is now the accepted terminology (e.g. Sharp,1994)
121	Para 2	Replace last sentence with	Such an attempt is the subject of the remainder of this study which also includes in Chapter 12, the construction of an appropriate quality framework based on the issues discussed here.
175	Figure 20	Delete / Replace with	Diagram below
179	Para 1 line 1	Delete sentence 2	
180	Section 2.4.3.1 List point 3	Delete/ Replace with	double headed arrow crowsfoot
180	Section 2.4.3.1 List point 4	Delete	or arrowhead
180	Section 2.4.3.1 List point 5	Delete/ Replace with	arrowheads barred crowsfeet
185	Para 3 line 6	Delete/ Replace with	in practice in this development
221	End of para 1	Insert	The development of the INTECoM framework has also raised a number of interesting issues for research, which are discussed in more detail from page 236.
229	Para 2 line 3	Delete	prescriptive
284	Fact Type 6 Example	Delete last 3 sentences	

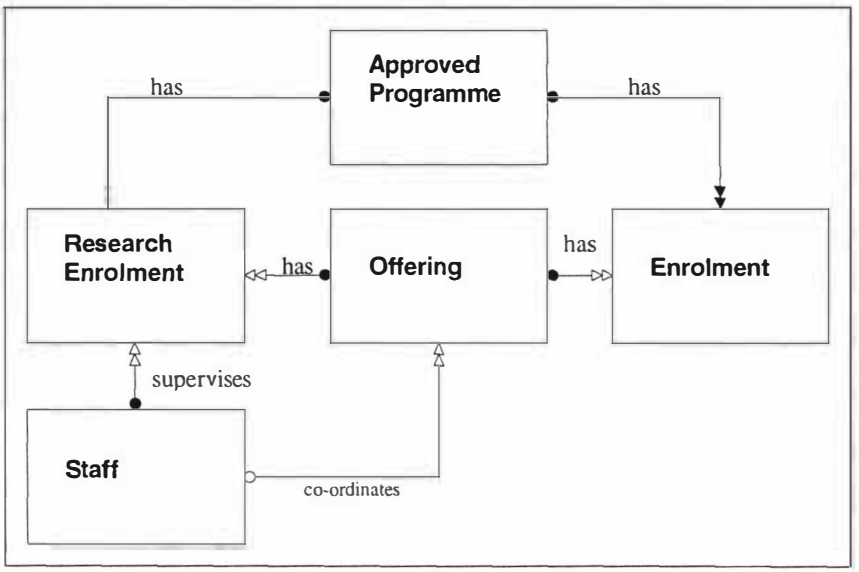

Figure 20 One solution to the Research Supervisor Problem

Abstract

Conceptual data models, a fundamental component of information systems development, traditionally play two essential roles, as communication tools and database design blueprints. However, despite their importance to the success of information systems, and a considerable amount of research effort, no definitive method for constructing them has yet been described. Entity-Relationship (E-R) Modelling, accepted as a *de facto* standard for a number of years, has been increasingly criticised. A number of alternatives have been proposed and some, such as Object-Oriented modelling, have gradually been accepted by the practitioner community. Nevertheless, effective conceptual data modelling continues to be recognised as a difficult activity, both to teach and to practice.

This study investigates conceptual data modelling in the context of the relational database development process. However, rather than specifying a new method or exploring the efficacy of existing ones, it focuses on the nature of the activity itself. The construction of a conceptual data model encompasses both the analysis and design stages of systems development. Some fundamental differences in modeller behaviour, required by these activities, are explored. The disparate purposes of a conceptual data model are also investigated and the effectiveness of using the same modelling method in both stages, and for both purposes, is questioned. To explore the suitability of current methods to specific development activities, the procedures inherent in the use of two conceptual data modelling approaches, E-R Modelling and the Natural Language Information Analysis Method (NIAM) are also investigated.

The result of this exploration is the recognition that the methods have exclusive strengths and that it is more productive to view them as complementary rather than competing. Consequently, a database design framework, INTECoM, is constructed in which the two methods are integrated and matched to the activities for which they are most suited. The framework is supplemented by a new technique, NaLER, to facilitate communication in the design stage. The soundness and viability of this theoretical framework is examined through its use on a small development and the implications of the adoption of INTECoM on both education and practice are considered.

Preface

> *"The inherent structure of the E-R approach can be easily explained using natural language. For this reason, the 5 -10 year time frame will bring the combining of the E-R approach with the linguistic approach. Future modeling scenarios will involve a natural language interface, where the E-R model underlies with an interpreter and is invisible to the user. (Hawryszkiewycz, 1987 p.471)*

Hirschheim *et al.* (1995) comment on the large amount of research that has been amassed in the fields of "information systems development, in general, and data modelling in particular". They continue,

> "A wealth of research in these fields has produced an astonishing array of empirical results and practical insights, conceptual and terminological diversity and confusion and a large suite of tools and methods. But as many researchers and practitioners alike feel, these form an isolated, disjoint and often contradictory amalgam of knowledge" (p.xi).

This researcher concurs with this view and acknowledges their contribution in highlighting the diversity and confusion, which undoubtedly exists. While this study does not include significant discussion of their ideas, their informed analysis and incisive comments underlie many of its observations. Via, a close analysis of the literature and a synthesis of existing knowledge within the area of "fact-based data modelling" (*ibid.* p.28), this study attempts to "to shed light on similarities where they exist and to discuss possible directions for improvement" (*ibid.* p.xi).

This research is motivated by the belief that the nature of the conceptual data model has changed fundamentally over the last twenty years and that both the nature of these changes, and the purpose of the conceptual model, are not always clearly understood. Practitioners and academics alike, provide many independent definitions relevant to their specific viewpoint or purpose, with few attempts to provide a holistic description. Indeed, the analogy of six blind individuals attempting to construct an accurate image of an elephant by discretely describing its constituent parts may well be appropriate (Pletch, 1989). Additionally, the variety of research methods that has been, and continues to be, used, result in research findings that are often perplexingly

contradictory, untimely, and notable for their inability to inform practice. Rather than adding more experimental work to this mélange, this study focuses on, and attempts to clarify, some of the issues arising from this fragmented understanding. The principles underlying the conceptual data modelling process are exposed and existing tools, techniques and methods are investigated. The purpose of this critical examination is to explore whether integration can improve both the process of conceptual data modelling and the appropriateness and quality of the outputs.

The philosophical basis for this research is thus broadly interpretivist. Interpretivism, which has recently "...emerged as an important strand in information systems research" (Klein & Myers, 1998 p.2), has been adopted by a number of IS researchers (e.g. Walsham, 1993; Boland, 1991). Klein and Myers (1998) comment that,

> "IS research can be classified as interpretive if it is assumed that our knowledge of reality is gained only through social constructions such as language, consciousness, shared meanings, documents, tools and other artifacts. Interpretive research does not predefine dependant and independent variables but focuses on the complexity of human sense making"(p.5)

Likewise, Walsham (1993) describes interpretivism as "an epistemological position concerned with approaches to the understanding of reality and asserting that all such knowledge is necessarily a social construction and thus subjective" (p.5). In discussing the use of interpretive methods for case study research, he later comments on the interrelationship between epistemology and research methods,

> "If one adopts a positivist epistemological stance, then statistical generalisability is the key goal. However, from an interpretivist position, the validity of an extrapolation from an individual case or cases depends not on the representativeness of such cases in a statistical sense, but on the plausibility and cogency of the logical reasoning used in describing the results from the cases, and in drawing conclusions from them" (*ibid.* p.15)

Although this study does not involve the construction of case studies, it does rely on the plausibility and cogency of its logical reasoning to make sense of a particular set of IS development activities. It makes no claim to provide empirical evidence of the validity of the conclusions that are drawn but lays down threads of discrete arguments that are synthesised into a credible scenario. In so doing, this researcher has been influenced by, and taken cognisance of the arguments of researchers such as Lyytinen and Klein (1985), Galliers (1993), Ciborra (1997), Walsham (1995), and Klein and Myers (1999).

Ciborra (1997), in particular, has identified a crisis in IS research which he suggests stems from the general adoption of the "paradigm of the natural sciences and the relevant methodologies of measurements, normalisation and calculation."(p.1551). He continues,

> "What I am concerned with here is something subtly pervasive: it is for example, that in order to show that structured methodologies are a failure or plainly not used, one has to adopt a structured scientific method to empirically measure the phenomenon, in order to be credible and legitimate; and even then, being methodologies at the core of our discipline, these empirically measured facts still get to be dismissed" (*ibid.* p.1552).

This, he argues, provides a disservice to the IS community as "we tend to forget the role of human choice behind the technical artefacts, and study the user side of information systems by adopting the methods of the natural sciences" (*ibid.* p.1552). In response to this tendency, some researchers have looked instead to the social science disciplines for more appropriate methods (e.g. Walsham 1995; Klein & Myers, 1999). Galliers (1993) has suggested that the mode of subjective/argumentative research, which is the primary form adopted in this thesis, is appropriate for investigating methodological issues and categorises it as a post-positivist, interpretive approach.

Most of the more recent research effort, in the conceptual data modelling area, has concentrated on positivist, empirical studies that either compare and refine various qualities of different modelling formalisms or observe the behaviour exhibited by modellers with differing levels of expertise. However, perhaps because of the difficulty in designing pertinent experiments, little attention has been paid to the "*ways of working*" (Bronts *et al.*, 1995 p.214) inherent in the different formalisms. No studies were discovered that investigated whether the different behaviour required of the modeller by different formalisms had any effect on the quality of the final product. Likewise, there has been little, if any, consideration given to the data modelling requirements of the different stages of the information systems life cycle. Consequently, the appropriateness, of different techniques for the various tasks required by these stages, does not appear to have been adequately examined either.

This study then, is not interested in attempting to re-assess whether one formalism is more expressive, useable or comprehensible than another. Instead, it explores the wide range of functions ascribed to the conceptual data model, and focuses on the different

working practices and behaviours inherent in different conceptual data modelling formalisms. It then investigates the different 'ways of working' required by the activities of analysis and design and explores the question of whether some formalisms are inherently more suited to certain stages of information systems development than others.

In order to establish a base from which a useful integration can be derived it is necessary to also explore a number of related areas. Therefore, this study also examines various issues including; the history of conceptual data modelling; pedagogical issues; and methods that have previously been used to evaluate the quality and effectiveness of conceptual data models. A more detailed look at the processes involved in developing a conceptual data model with both Entity-Relationship (E-R) and NIAM (Natural Language Information Analysis Method) techniques is also undertaken. The end result of this critical and hermeneutic examination is the delineation of a framework within which these particular techniques can be used to greater advantage than the use of either of them alone.

The apparent contradiction in investigating the use of data modelling tools, classified as objectivist by Klein and Hirschheim (1987), to undertake an activity, data modelling, classified by them as subjectivist, is recognised. However, this study focuses on current education and practice and it is useful to accept and work within this apparent paradox. As Weber (1997) points out the ontological assumptions underlying the data modelling activity comes from the modellers themselves rather than from the data modelling grammars they choose to use. This study also broadly concurs with Kent's (1978) observation that,

> "*[A]*t bottom we come to this duality. In an absolute sense there is no singular objective reality. But we can share a common enough view of it for most of our working purposes so that reality does appear to be objective and stable" (p.203).

Following a general introduction, and a working definition of some of the more overloaded terms in the first two chapters, Chapter 3 provides a historical perspective on the development of the data modelling activity. In so doing, the chapter explores the alteration in the purposes and use of a conceptual data model and raises some fundamental questions regarding the activity itself. Chapter 4 explores these issues in

more detail and investigates their implications on both practice and education. The process of constructing an E-R Model is discussed in Chapter 5 and the predominance of the Chen (1976) E-R Model, widely assumed by many academic writers, is challenged. In particular, the descriptive, creative nature of the process of E-R modelling and its appropriateness for analytical activity is explored. An alternative method, NIAM Conceptual Schema Design Procedure (NIAM-CSDP), is described in Chapter 6, with a continuing investigative emphasis on the procedure by which a model is created.

Having thus established a basis for such a discussion, Chapter 7 provides a comparison of the 'ways of working' of the two techniques. There is no attempt to establish any superiority of one method over another. Instead, the emphasis is on matching the different approaches to appropriate stages of the information systems development life cycle. In Chapter 8 the perspective of the analysis shifts slightly, to investigate the means by which the quality of conceptual models has been evaluated. This investigation confirms the existence of diverse definitions of a conceptual data model and highlights a number of issues that spring from this lack of consensus. It also shows that much of the previous research has sought to establish one modelling facility as the 'best', to the exclusion of all others. This study takes the position that the search for a 'holy grail' of data modelling, is inappropriate, wasteful and dangerous and has diverted attention away from both an investigation of the nature of the activities required by database development and on the construction of tools to match those needs.

Consequently, Chapter 9 proposes a framework, INTECoM in which elements of the NIAM and E-R approaches are matched to appropriate needs and used productively together. Chapter 10 provides a new technique, NaLER for extracting NIAM-like sentences from E-R models. It justifies the need for such a technique primarily as a means of constructing a formal, natural language view of the final design model and facilitating the audit of the conceptual data modelling process, but also to assist in model interpretation for both non-technical users and inexperienced modellers.

A small development to demonstrate the viability of the INTECoM framework was undertaken and is described in Chapter 11, while the issues of quality evaluation related to the use of the framework are discussed in Chapter 12. The final version of INTECoM

is defined in Chapter 13 and the study concludes, in Chapters 14 and 15, by looking at the implications inherent in the adoption of the INTECoM framework, for both the education of future data modellers and the practice of data modelling. The implications of using the NaLER technique is also discussed and some potential criticisms are addressed. A number of limitations of this work are highlighted and the possibilities for future research are delineated.

This thesis then, focuses on theory building rather than theory testing, constructing a new theory, in the form of an integrated framework, in which each element is justified by reference to, or inference from, the relevant literature. By taking a holistic view of the development of the data modelling process, it exposes and explores questions that have been largely overlooked by previous researchers. A potential criticism of this work could be that no empirical studies have been undertaken to assess the efficacy of the proposed framework. Apart from the concentration on theory building, there are several reasons for this. Firstly, the primary techniques within the framework, ER Modelling and Object-Role Modelling, are already well tested, used in the database community and considered to be effective. Secondly, the nature of the proposed framework is such that for a case study to hold any real significance it would need to be tested on a medium to large 'real-world' project. In addition, while such a test would undoubtedly be profitable, the scale of such a test puts it beyond the scope of this research. Instead, a small development undertaken by the researcher is described in detail, as a worked example, to demonstrate the overall structure of the framework, the inter-relationships that exist between the various elements and how the framework can be successfully instantiated.

There is one very clear and deliberate omission from this work and that is a detailed consideration of the object-oriented (O-O) approach to system development. While this omission is intentional and stems from a number of factors, O-O techniques have not been ignored. However, while the use of object-oriented techniques is undoubtedly increasing, as yet no standard process, whether *de facto* or actual, exists. The Unified Modelling Language (UML), currently supported commercially by Rational Software Corporation, brings together the major O-O methods that have been developed over the last ten years Booch (Booch, 1991, 1995), Object Modelling Technique (OMT) (Blaha *et al.* 1988; Rumbaugh *et al.*, 1991; Blaha & Premerlani, 1997) and Objectory (Jacobsen

et al., 1992). UML provides a standard notation and development approach by bringing together the design strengths of Booch, the analysis strengths of OMT and the strong behavioural analysis of use cases (Quatrani, 1998). However, while the eventual acceptance of UML, as a *de facto* standard is probable, it is not yet recognised as such.

A more important consideration is that neither UML nor its predecessors provide a proven means of designing relational data structures. Textbooks describing the object-oriented approach to system development tend to emphasise the process and software aspects of development and either ignore the need for database design (e.g. Quatrani, 1998) or treat it as a required but additional technique (e.g. Larman, 1998). Indeed, as Quatrani (1998) states, "the Rational Objectory Process *[is]* an extensive set of guidelines that addresses the technical and organisational aspects of **software**[1] development" (p.8). The issues considered here are very specifically related to the development of relational databases and the detailed exploration of methods that do not seek to provide a means to design such structures is thus not relevant. Of all the O-O methods, OMT had placed the greatest emphasis on relational database design and the general approach to system development recommended in both UML and OMT is considered in Chapter 5. However, it is argued that while both the notation and the working language of these formalisms may differ, in terms of their approach to data identification and structuring, their 'way or working'[2] is essentially no different from the approach required by traditional E-R Modelling. Indeed, Eaglestone and Ridley (1998) specifically state,

> " Data analysis methods provide systematic processes by which conceptual models are derived. One such method, more commonly associated with relational database design, is entity-relationship analysis [Chen 76]. This is not inherently linked to relational databases and is in fact the sort of process we would wish to undertake to design an object database" (p.276).

This research takes the view that the requirement to build sound data structures is likely to continue to be important for some time and that the need to formally analyse and record user data needs and to then transform them to effective relational databases, will thus remain. While the specific techniques that are used to meet these requirements may

---

[1] Emphasis added

[2] The meaning and importance of a method's 'way of working is considered fully in Chapter 5.

well change, the fundamental principles that underlie them are likely to remain constant. It is thus the adoption of the principles of the INTECoM framework that is advocated by this thesis rather than any specific data modelling formalisms with which it is instantiated.

Acknowledgements

My initial thanks go to my ex-colleagues at the Statistics Division of the Inland Revenue in the UK. In particular, Dr Ron James who first inspired me to a deep interest in data modelling and Dave Boutwood, who not only taught me the value of constructive argument but also gave me plenty of practice.

For more specific help, I must thank my two supervisors at Massey University: Dr Daniela Mehandjiska-Stavreva, now of Bond University, Australia for her support and encouragement in bleak times and Professor Jon Patrick, now of the University of Sydney, for the challenging intellectual debate he provided and particularly for his assistance in formally defining the NaLER language.

In addition, Wen van Kersbergen, from the Amsterdam School of Business, gave me his invaluable guidance on the practical use of the NIAM-CSDP and Info*Modeler*™ and Dr Steve Hitchman provided me with many useful insights and criticisms, from both his academic and professional experience.

All my colleagues in the Department of Information Systems, particularly Chris Freyberg, Mike Ryder and Peter Blakey have given me both practical and moral support. Lastly, my heartfelt thanks to all the students on whom I have tested out my theories and explored my ideas; for their feedback, their enthusiastic participation and their patience.

Publications

The following refereed papers have been directly based on work in this thesis.

Atkins, C.F.(1996): Prescription or Description: Some Observations on the Conceptual Modelling Process, in PURVIS, M.(ed.), Proceedings *of Software Engineering: Education and Practice Conference,* Dunedin, New Zealand, January: 34-41.

Atkins, C.F. and Patrick, J.D.(1998): NaLER: A Natural Language Method for Interpreting E-R Models, in PURVIS, M. (ed.), Proceedings *of Software Engineering: Education and Practice Conference,* Dunedin New Zealand, January: 2-9.

Atkins, C.F. and Patrick, J.D.(2000): NaLER: A Natural Language Method for Interpreting E-R Models, Campus *Wide Information Systems,* (forthcoming).

Contents

List of Figures

List of Tables