SEMANTIC INTEGRITY IN DATA WAREHOUSING: A framework for understanding.

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

Masters of Business Studies

in

Information Systems

at Massey University, Palmerston North
New Zealand.

Jennifer Jane Sampson

2001

Abstract

Data modelling has gathered an increasing amount of attention by data warehouse developers as they come to realise that important implementation decisions such as data integrity, performance and meta data management, depend on the quality of the underlying data model. Not all organisations model their data but where they do, Entity-Relationship (E-R) modelling, or more correctly relational modelling, has been widely used. An alternative, dimensional modelling, has been gaining acceptance in recent years and adopted by many practitioners. Consequently, there is much debate over which form of modelling is the most appropriate and effective. However, the dimensional model is in fact based on the relational model and the two models are not so different that a debate is necessary. Perhaps, the real focus should be on how to abstract meaning out of the data model.

This research explores the importance of semantic integrity during data warehouse design and its impact on the successful use of the implemented warehouse. This has been achieved through a detailed case study. Consequently, a conceptual framework for describing semantic integrity has been developed. The purpose of the framework is to provide a theoretical basis for explaining how a data model is interpreted through the meaning levels of understanding, connotation and generation, and also how a data model is created from an existing meaning structure by intention, generation and action.

The result of this exploration is the recognition that the implementation of a data warehouse may not assist with providing a detailed understanding of the semantic content of a data warehouse.

Acknowledgements

Thanks to the case study participants for their enlightening discussions throughout the interviewing process, I really appreciated the time they dedicated to the interviews.

I would also like to thank Dianne Wheeler for applying the case study guidelines from this research to prepare the critical appraisal guidelines for single case study research.

I am extremely grateful to my supervisor Dr Clare Atkins for her guidance and enthusiasm during this research. I would like to thank Clare for all the fun times we had putting together this research, I shall miss our weekly meetings. I also appreciate the morale support you have given me over the past two years.

Most of all I would like to thank my family for their support throughout this research. A special thanks to my mother for all the times you have looked after Heather and Harriet, I shall be forever grateful. Thanks also to my two daughters for their patience and loving smiles.

Contents

1 INTRODUCTION 1

2 DETERMINING SEMANTIC INTEGRITY: A FRAMEWORK FOR UNDERSTANDING. 9

INFORMATION AND MEANING 9
LEVELS OF MEANING 20
INTERSUBJECTIVE MEANING IN DATA MODELLING 25
DATA MODEL QUALITY: THE WIDER CONTEXT 31
SUMMARY 35

3 DATA MODELLING FOR THE DATA WAREHOUSE 37

INTRODUCTION 37
THE DATA WAREHOUSE 39
DATA MODELLING FOR THE DATA WAREHOUSE 40
THE PROCESS OF DATA WAREHOUSE DESIGN 47
SUMMARY 51

4 THE RESEARCH PROCESS 53

INTRODUCTION 53
RATIONALE 55
FORMULATION OF THE RESEARCH QUESTION 55
DATA COLLECTION METHODS 56
QUALITY CONTROL MEASURES 57
SUMMARY 58

5 PRACTICAL GUIDELINES FOR SINGLE CASE STUDY RESEARCH 59

INTRODUCTION 59
PREVIOUS RESEARCH 59
DETERMINING PRACTICAL GUIDELINES FOR CASE STUDY RESEARCH 61
PRACTICAL GUIDELINES FOR CASE STUDY RESEARCH 63
SUMMARY 77

6 THE DRESS REHEARSAL 79

INTRODUCTION 79
THE PILOT CASE STUDY 79
FORMAL CASE STUDY CRITERIA 84
FUTURE RESEARCH - PROJECT CRITERIA 86
SUMMARY 88

7 INVESTIGATING SEMANTIC INTEGRITY: A DETAILED CASE STUDY 89

THE CASE STUDY SETTING 89
PAINTING THE PICTURE 92
GENERATION OF MEANING FROM A DATA MODEL 93
PRODUCING A DATA MODEL FROM MEANING 112
SUMMARY 122

8 STRATEGIES FOR ACHIEVING SEMANTIC INTEGRITY 125

INTRODUCTION 125
GENERATING MEANING FROM A DATA MODEL 125
PRODUCING A DATA MODEL FROM MEANING: GENERAL STRATEGIES 134
SUMMARY 135

9 CONCLUSIONS AND FUTURE RESEARCH 137

REFLECTION 137
SHORTCOMINGS 140
FUTURE WORK 141

SHORTCOMINGS 140
FUTURE WORK 141

GLOSSARY 153

APPENDICES 155

APPENDIX 1: PILOT CASE STUDY QUESTIONS 157
APPENDIX 2: TECHNICAL PILOT CASE STUDY QUESTIONS 161
APPENDIX 3: USER CASE STUDY QUESTIONS 163
APPENDIX 4: TECHNICAL CASE STUDY QUESTIONS 167
APPENDIX 5: CRITICAL APPRAISAL GUILDEINES FOR SINGLE CASE STUDY RESEARCH 171
APPENDIX 6: Q.S.R. NUD*IST4 INDEX STRUCTURE 177

List of Figures

Figure 1: The Conventional View on the Knowledge Hierarchy (Tuomi, 1999) ... 11
Figure 2: Information in the Interpersonal Process (Tuomi, 1999, p.112) 12
Figure 3: Structuring principles of ICMs 16
Figure 4: Levels of meaning (Mingers, 1995, p.299) 23
Figure 5: Intersubjective Meaning in Data Modelling 26
Figure 6: Components for Understanding Data and Metadata Quality in a Data Warehouse (Shanks and Darke, 1998a, p.125) 32
Figure 7: INTECoM - Quality Framework (Atkins, 2000) 34
Figure 8: Four models for warehouse design (Winsberg, 1998, p.510). 41
Figure 9: The Research Process (adapted from Shanks *et al.*, 1997) 54
Figure 10: Initial MIS Architecture (Lawrence, 1997a) 91
Figure 11: Current MIS Architecture. 91
Figure 12: Data Warehouse Approach (adapted from Lawrence, 1997a) 92

List of Tables

Table 1: Normal form description for water (Putman, 1975, p.191) 14
Table 2: Image Schemas for a Data Model 19
Table 3: Levels of Meaning 24
Table 4: Nine guidelines for evaluating papers (Greenhalgh, 1997) 61
Table 5: Guidelines for Case Study Research 63
Table 6: Pilot Case Study Findings 82
Table 7: Four classifications of data warehouse projects. 87
Table 8: Stakeholder Data Requirements (adapted from: Mair, 1999). 90
Table 9: Inhibiting Factors for Understanding 95
Table 10: Inhibiting factors for Connotation 105
Table 11: Inhibiting Factors for Intention 109
Table 12: Inhibiting Factors for Generation 114
Table 13: Strategies for Understanding 130
Table 14: Strategies for Connotation 133

1 Introduction

> *"Our knowledge of the existence of cells seems secure, as secure as any knowledge is likely to be. Nonetheless, it is human knowledge based on human understanding, not on any neutral, or God's-eye-view, understanding. There is no such thing as a neutral way to understand things. But as long as our human understanding remains stable, it is possible for our knowledge to be secure"* (Lakoff, 1987, p.300).

The use of semiotics for understanding data quality in IS has been discussed by a number of researchers (Benyon, 1997; Hirschheim *et al.*, 1995; Mingers, 1995; Shanks and Darke, 1998a; Shanks and Corbitt, 1999; Stamper, 1987). However, this research focuses on exploring the importance of semantic integrity, and applies a framework based on semiotics to describe intersubjective meaning in data modelling. The research of Mingers (1995) provided the groundwork for this research. He writes,

> "Computers process (transmit and transform) signs (data) and the information which they carry. In itself, this information is quite meaningless until it connects to the wider meaning systems within which human beings operate. What we call information systems are really only a part of human meaning systems in which signs and signals are continually produced and interpreted in an ongoing process of intersubjective communication" (*ibid.* p.303).

There has been little academic research which examines semantic integrity in the context of data warehousing, although data warehousing is a rapidly growing area of interest to many organisations. This research explores the problem of defining 'meaning' in a data model and the implications of this for data warehouse design. De Carteret & Vidgen (1995) describe this when they comment, "The meaning is not entirely in the data model and it is not entirely in the situation being modelled - it lies somewhere between the two and cannot be located precisely" (p.373). The framework proposed in this research is useful as an initial description of this grey area.

Mingers (1995) comments on the importance of semantic and pragmatic meaning,

> "For practical IS development, empirics and syntactics are necessary, but it is the semantic and pragmatic aspects of information, where signs gain meaning and are used, that is crucial" (p.286).

Atkins (2000) notes that in the pre-relational environment of the ANSI/X3/SPARC (1975) report, 'users' were either computer programs or computer programmers. Because of the nature of such users it was unnecessary to "undertake extensive validation of whether the representation of the data structure that the designer had created, matched the users' own view of the data structure" (p.41). Often was the case that if the data structures did not successfully support the user requirements, the requirements were changed rather than the database structure. Today end users tend to be "people with relatively few technical skills but extensive enterprise knowledge" these people have "both the opportunity and the desire to directly access the data of interest to them" (*ibid.*p.42). Therefore, a data modelling approach must provide an adequate communication device for explaining to human users the semantic content of the model.

Additionally, this research area is important because implementation decisions such as data integrity, performance and metadata management, depend on the quality of the underlying data model (Devlin, 1997; Inmon, 1993; Kimball, 1996; Mattison, 1996, Silverston *et al.*, 1997).

As data modelling is concerned with the representation of knowledge, "a philosophical background on human inquiry and the nature of knowledge is pertinent for understanding the problems of data modelling" (Hirschheim *et al.,* 1995, p.145). Hirschheim *et al,* (1995) classify three paradigms of data modelling: functionalism, social relativism and neohumanism, however, they remark that research literature in IS continues to promote one paradigm, functionalism in information systems development and objectivism in data modelling.

They ask four questions of each paradigm: the ontological question (what is being modelled?); the epistemological question (why the result is valid?); the social context question (what is the relationship between the social world and the data modelling?); and the representation question (how is the result presented?). The case study undertaken in this research gathered information relating to these questions but focused on the epistemological question (why the result is valid).

Furthermore, they describe data model validation from these three perspectives, firstly from a functionalist epistemological stance, they comment,

> "valid data models can be built by applying proper observation and data collection methods to an object system, i.e. the application domain. ...its accuracy can be determined by checking how well it corresponds to the reality of the object system. By observing the deficiency of the application, one can infer the likely cause in the specification and correct it. In this way the data model can be tuned over time to improve its correspondence with reality" (*ibid* p.158).

Practitioners typically accept this objective approach, however such an approach may cause problems that become expensive to correct once the database is built. From a social relativist epistemological perspective, a data model "can be more or less accurate or more or less appropriate" (*ibid* p.162). Hirschheim *et al.* (1995) continue by suggesting three principles to guide practice, research and methods of data modelling from a social relativist epistemological stance:

> "(a) All data models have fundamental bias that can be traced to the contingent preunderstandings with which they were built.
>
> (b) To some extent, the bias can be made transparent through bracketing, a form of self-critical, reflective dialogue.
>
> (c) Bracketing must not be seen as a procedure to decide between fundamentally conflicting preconceptions. Therefore a hermeneutic approach to data modelling is very skeptical of the idea that bias can eventually be substantially reduced or even be eliminated by a process of evaluative elimination" (*ibid.* p.162).

Thirdly, Hirschheim *et al.* (1995) describe the epistemological perspective of neohumanist data modelling.

> "To be true, the implications of a data model must be 'warranted', that is to say that the fundamental perspective and simplifying assumptions which are inescapably built into any model must be legitimised through an informed consensus. From this it follows that the most appropriate data modelling must be informed by the widest possible participation" (*ibid.* p167).

While this classification of data modelling paradigms may be interesting, de Carteret & Vidgen (1995) argue that an interpretative approach, which recognises the benefits of both objective and subjective aspects, is more appropriate. However, some of the principles they suggest may be useful as input for developing strategies for semantic integrity.

The use of a data warehouse is dependent on the provision of information that is meaningful to the end users. Newcum (2000) comments from a pragmatic point of view "Quality is really only useful to business people who have to gather data to turn into information (and perhaps even into wisdom) so that they can make business decisions".

An important area for research is one which explores the problem of how different stakeholders interpret the information carried by the data warehouse. This research explores this problem and describes strategies to help achieve semantic integrity. This is important since one of the goals for data warehouse design is to develop a design data model that may be understood by the different stakeholders. Hirschheim *et al.* (1995) mention this when they comment, "Business data are such a standard set of signs which are expected to convey the same or at least similar meanings to a user community" (p.14).

Little formal research has been conducted to explore the importance of semantic integrity and its impact on the successful use of the implemented warehouse. However, Shanks and Darke (1998a) have proposed a framework for understanding data quality in a data warehouse (described further in chapter two).

There are many practitioner publications on the subject of data warehouse development, most of which cover data modelling to some degree. However, as Date (2000) points out the discussion is usually from a physical perspective promoting the dimensional model (or star schema). However, the activity of data model validation is generally not discussed.

The main purpose of this research is to explore the importance of semantic integrity during data warehouse design and its impact on the successful use of the implemented warehouse. This will be achieved through a detailed case study.

Propositions:

1. Semantic integrity is an important critical success factor in determining the effectiveness of a data warehousing project.

2. A 'good' data model is an important critical success factor in determining semantic integrity.

Semantics deals with the issue of 'meaning' that is, the relationship between signs and what they are supposed to represent (Stamper, 1987). Semantic quality can be described according to two concepts: structure and content (Shanks and Darke, 1998a). The structure (or metadata) refers to the representation of the stakeholder domain models using some language, for example the dimensional model. The goals for semantic quality according to the **structure** of the data warehouse are: *completeness* and *validity* (Lindland *et al,* 1994). Whereas the goals for semantic quality according to the **content** (the data) of the data warehouse are: *completeness* and *accuracy* (*ibid.* p.126). However, this research will focus on the importance of intersubjective meaning, and suggests two additional goals for semantic integrity *meaningfulness* and *comprehensibility*. 'Comprehensibility' may be appropriate in terms of both the structure and the content, however, 'meaningfulness' may be appropriate in terms of the content of the data warehouse. A framework is presented in chapter two incorporating Mingers (1995) levels of meaning, this represents the generation of meaning from a data model and the production of a data model from meaning.

The intellectual framework for this research is based on the underlying ontological, epistemological and methological beliefs. In the interpretive tradition the ontological[1] position of constructivist is taken. The constructivist position is that,

> "the domain of interest exists independently of any stakeholder, but that the cultural background and knowledge of the stakeholder influences the perception and subsequent representation of that domain. Therefore representations of any domain (that is, data or metadata) may be interpreted differently by stakeholders and are subject to negotiation among communities of stakeholders" (Shanks and Darke, 1998a, p.124).

The epistemological[2] position can be viewed as broadly interpretive "seeing the pursuit of meaning and understanding as subjective and knowledge as a social construction" (Walsham, 1993, p.21). The methological approach is an exploration of the importance of semantic integrity during data warehouse design, while the research method involves the use of a single case study. As there is very little research in data warehousing (Shanks *et al.* 1997*),* and there is a specific lack of

[1] Ontology refers to the nature (or theory) of reality.

[2] The belief about how knowledge is acquired.

research into the activity of data modelling for a data warehouse, Benbasat *et al.* (1987) would argue that a case study method is 'suitable', as the problem is one where "research and theory are at their early, formative stages" (p.369). A single case study is suitable for this research since the objective is to explore in **detail** the importance of semantic integrity during data warehouse design and its impact on the successful use of the implemented warehouse.

While a case study approach may be suitable, it is important to recognise the difficulties with finding and then gaining access to both appropriate projects and the relevant participants. Originally, the intention was to perform four case studies, however, because of the difficulties associated with finding appropriate projects, only one case study was performed. Ultimately, studying one project allowed a detailed analysis to be carried out, revealing inhibiting factors for both the generation of meaning from a data model and the production of a data model from meaning. Such a detailed analysis may not have been feasible if multiple case studies had been performed. However, while the data analysis undertaken was detailed, it was not sophisticated. Future research may involve undertaking further comparisons within the data and using multiple case studies. Nevertheless, this research has proved fruitful for providing strategies for achieving understanding of the physical data model for the particular organisation studied.

Apart from these problems, other problems may have resulted due to the choice of a case study method. For example, the researchers background may have influenced the data collection and data analysis. In addition, the integrity of this research relies on an objective interpretation of the actual events (Galliers, 1993).

Fundamental to this research is the use of a conceptual framework for describing semantic integrity. The purpose of the framework is to provide a theoretical basis for explaining how a data model is interpreted through the meaning levels of understanding, connotation and generation, and also how a data model is created from an existing meaning structure by intention, generation and action. These ideas and others relating to cognitive semantics (Lakoff, 1987) are discussed in chapter two. Furthermore, because there is little research on data modelling for the data warehouse, it was necessary to examine the existing literature. Date (2000) provides the most rigorous description of both logical and physical data modelling for the data warehouse, this is discussed in chapter three.

This research has also involved developing guidelines for single case study research. These guidelines are the quality control measures for this research (refer chapter five) and were necessary as no existing unified list of criteria for single case study research was found. A pilot study case study was undertaken which provided a low risk environment for verifying the research questions. This was an important activity which generated change in the research design and provided conceptual clarification (refer to chapter six).

The framework presented in chapter two also serves as the structure for describing the case study findings in chapter seven. For each meaning level, inhibiting factors are described based on the case study findings. Finally, general and specific strategies for semantic integrity are suggested in chapter eight.

2 Determining Semantic Integrity: A framework for understanding.

"The highway authorities are curiously reluctant to impart much in the way of useful information, like where you are or what road you are on. This is all the more strange when you consider that they are only too happy to provide all kinds of peripheral facts – NOW ENTERING BUBB COUNTY SOIL CONSERVATION DISTRICT, NATIONAL SPRAT HATCHERY 5 MILES, NO PARKING WED 3A.M. TO 6A.M., DANGER LOW FLYING GEESE, NOW LEAVING BUBB COUNTY SOIL CONSERVATION DISTRICT" (Bryson, 1989, p.40).

Information and Meaning

There is no consensus within the IS field over the nature of information. However, the most common, traditional view of information was determined by Lewis (1991) through a survey of introductory IS texts as "data that had been processed in some way to make it useful for decision makers" (quoted in Mingers, 1995, p.285). This implies that data is objective and that information "can be objectively defined relative to a particular task or decision" (*ibid.* p.285). The alternative view is that information is subjective because different people will create different information from the same data due to their differing values, beliefs and expectations (Lewis, 1993: quoted in Mingers, 1995). However, Mingers (1995) argues that both views have significant weaknesses when he comments,

> "information is objective, but ultimately inaccessible to humans, who exclusively inhabit a world of meaning. Meaning is essentially *intersubjective* - that is, it is based on a shared consensual understanding. The implication is that information is only a part of what we understand by IS and that attention needs to be focused on the *meaning systems* within which *information systems* reside" (*ibid.* p.286).

Similarly, but in the context of IS development de Carteret & Vidgen (1995) argue that it is "not appropriate to consider the process as objective (IS development mirrors organisational reality) or as subjective (organisational reality is created through IS development)"(p.372). They claim that both objective and subjective aspects are present in all aspects of the IS development process at the same time.

Mingers (1995) clearly defines the basic concepts of the data, information and meaning and Mingers (1995b) also undertook an evaluation of information theories by analysing them according to four criteria. He concluded that Dretske's (1981) analysis of information and meaning were not only the most successful in terms of these criteria[3], but also the most suitable basis from which to develop theories of information at the semantic and pragmatic levels.

The main implication of this work was to provide a framework within which different world views may be located and integrated more effectively. Of particular relevance to this research is the focus on semantic and pragmatic dimensions of information, where he brings together the research of Maturana (1975) and Habermas (1984) into Dretske's (1981) framework of semantic information. Data, information and meaning are defined through an analysis of signs, symbols and utterances. Data is defined by Mingers (1995), as "a collection of signs brought together because they are considered relevant to some purposeful activity" (p.293), in the information systems domain, "data will usually be symbolic (numeric, linguistic or graphic) utterances, produced in the system for a particular purpose" (*ibid.* p293). Based on Dretske's (1981) argument 'information' is said to be the propositional content of a sign where the occurrence of a sign implies the information about the phenomena. Whereas 'meaning' is *generated* from information by interpreters through a process of digitialisation that extracts only some of the information available. Meaning is thus generated from information and leads on to action (Mingers, 1995). These definitions are appropriate for this research.

However, Tuomi (1999) disagrees with this description of data and information and argues that "the traditional hierarchy of data, information, and knowledge needs to be reconsidered if we want to develop information system support for knowledge management and organisational memory" (p.104). The implications of the reversed hierarchy are to promote the importance of tacit[4] and socially shared components, "If

[3] Theory evaluation criteria (Mingers, 1995, p.288):

(a) Generality of conception - the extent to which the theory provided a comprehensive and coherent description of information and meaning;

(b) Adequacy of concepts as a base for information systems in both theory and practice;

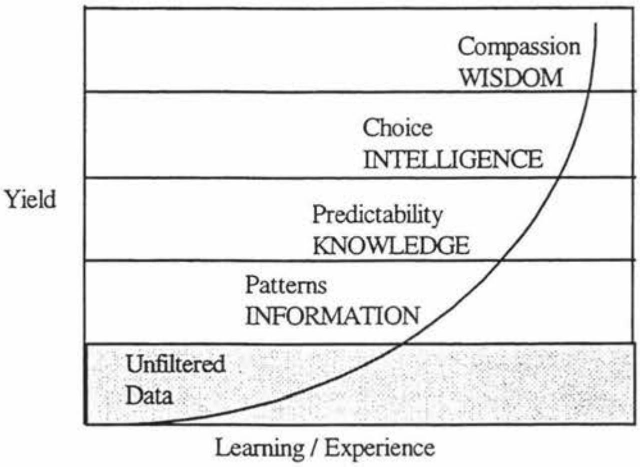
(c) Degree of fit with other theoretical and philosophical knowledge of other disciplines;

(d) Extent to which the theory corresponds to common-sense usage of the terms 'information' and 'meaning'.

[4] Tacit knowledge is personal, context-specific, and therefore hard to formalise and communicate (Nonaka-Takeuchi 1995: quoted in Tuomi, 1995, p.110).

the design principles cannot address the tacit component, it cannot tell us where and how much we should invest in the explication of knowledge" (*ibid.* p.113).

Similarly, Hirschheim *et al.* (1995) disagree with the traditional view that the meaning of an invariance is identical to the behaviour it results in. Instead, they believe that, by themselves invariances have no intrinsic meaning. Clearly this is because "interpretation is a creative act and no two interpretations are ever quite the same. Hence *meanings* are in the eye of some human beholder(s)"(p.13). They continue "Meaning is related to human understanding: through meaning we make sense of our feelings, thoughts and the world around us" (*ibid.*p.13).



Yield = intellectual dividends per effort invested

Figure 1: The Conventional View on the Knowledge Hierarchy (Tuomi, 1999)

Indeed, the perceived hierarchy order is dictated by the underlying ontological assumptions. It is beyond the scope of this research to investigate this debate in detail, however, it is important to recognise that the potential for both views exists. The conventional view on the knowledge hierarchy is shown at Figure 1.

Tuomi (1999) says there are several variations of the knowledge hierarchy, but data is generally seen as facts, which when structured become information, and in turn become knowledge when put into context or meaning is added. He argues that "the hierarchy of data-information-knowledge should be turned around. Data emerge last - only after knowledge and information are available. There are no "isolated pieces of simple facts" unless someone has created them using his or her knowledge" (p.107). This view of how data about an organisation is interpreted, is shown at Figure 2. According to this view a person articulates knowledge using the language(s) and conceptual systems available. For example when developing a database the articulated knowledge is represented using a predefined conceptual schema, this is then interpreted by someone else who tries to recover the potential meaning in the data. Tuomi (1999) comments that, "the success of this sense making depends on a sense maker's stock of tacit knowledge" (*ibid.*p.112). The original articulator and the sense maker need to have an interwoven meaning structure, that is "they have to share some world where the data can make sense" (*ibid.*p.112). Therefore "the sense maker must approach the data as meaningful data, that is as data intended to mean something" (*ibid.*p.112).

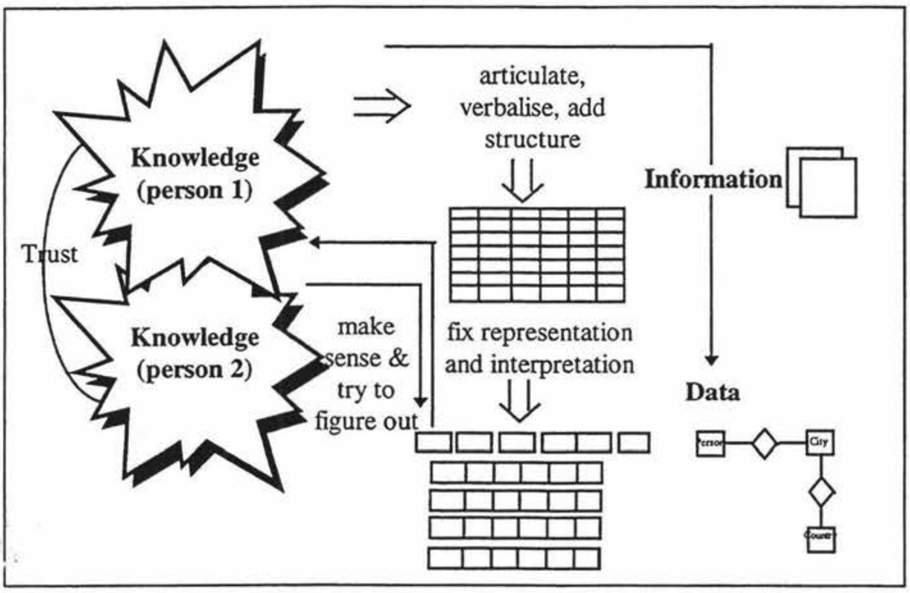

Knowledge (person 1)
articulate, verbalise, add structure
Information
Trust
make sense & try to figure out
fix representation and interpretation
Knowledge (person 2)
Data
Person
City
Country

Figure 2: Information in the Interpersonal Process (Tuomi, 1999, p.112)

On the contrary, Mingers (1995) says that meaning is derived from information. More specifically he says, "*information* is converted into (inter)-subjective *meaning* through a process of digitalization" (p.294). He continues "a message may carry information but have no meaning for a particular person who does not understand the language since they are unable to digitalize …while information must always be true, the meaning or belief we generate from information may be false - we must be mistaken"(*ibid.* p.294).

Unlike other researchers Mingers (1995) also relates 'what' and 'how' information is transmitted from a source to a receiver, writing

> "The amount of information that can be carried is calculated for both source and receiver. The question is how much of the information at the receiver is caused by the source? If there is complete transmission, it means that every state of the source is linked with every state of the

> receiver and vice versa. In practice this situation is unlikely. The receiver will be affected by things other than the source (noise), and not all of the information from the source will affect the receiver (equivocation)" (p.347-289).

(Therefore, the amount of information carried by or generated by the data warehouse is dependent on how 'successfully' the data is transferred from existing databases, to the data warehouse, to the end users.)

These issues raise the question of whether information on its own can ever be wholly semantically accurate or whether it is the meaning(s) that people derive from the information that defines the semantic integrity of the data warehouse. Similarly, Lakoff (1987) writes, "Meaning involves what is meaningful to us. Nothing is meaningful in itself. Meaningfulness derives from the experience of functioning as a being of a certain sort in an environment of a certain sort" (p.292). Determining semantic integrity for the data warehouse involves a subjective assessment of the 'meaning' stakeholders derive from the information, which cannot be achieved using a quantitative technique. Furthermore, a data modelling activity, can never, fully guarantee the semantic integrity of a data warehouse, although, the process of undergoing a data modelling exercise may be an essential factor when determining semantic accuracy. Because of this, a prototyping approach is often an appropriate approach to data warehouse development (refer pilot case study, chapter 5). Shanks and Darke (1998a) also agree when they suggest that prototype development with stakeholder involvement is one way to achieve *completeness*[5], (one of the goals[6] for semantic metadata quality).

Hirschheim *et al.* (1995) also compare data and information, they comment that, "*data* correspond to stating something (be it true or not) while *information* corresponds to speech acts which convey intentions" (p.14). They continue, "items of information are meanings that are intended to influence people in some way" (*ibid.* p.14). Like Mingers (1995) they discuss Habermas' *Theory of Communicative Action* in particular, how information corresponds to a speech act which makes an

[5] Shanks and Darke (1998a) define completeness as "the degree to which the data warehouse structure represents each of the concepts in the stakeholder's conceptualisations of the domain of interest" (p.125).

[6] The two goals for semantic metadata quality are completeness and validity (Lindland *et al.*, 1994: quoted in Shanks and Darke, 1998a). They also define quality in the data warehouse content, described later.

explicit truth claim. Habermas' (1984) four types of speech acts are: constative (to get someone to accept something as true), imperative (to get someone to do something), regulative (to appeal to others to obey accepted social norms) expressive (to express how one feels or thinks). Hirschheim *et al.* (1995) incorporated Habermas' (1984) validity claims to define knowledge,

> "If beliefs are stated about a subject with legitimate claims to truth or correctness, they are called *knowledge*. The difference between opinions or beliefs and knowledge is that the reasons or grounds for supporting the truth claims of knowledge have been approved by some qualified elite and therefore at the present time are taken to be beyond questioning for practical purposes (over time knowledge changes)" (*ibid.* p.14)

Putman (1975) too, argued that the traditional concept of meaning was one which rested on a false theory. Unlike the objectivist approach, he noted, that meaning is not in the mind. The assumptions based on the traditional theory of meaning were: firstly that, "knowing the meaning of a term is just a matter of being in a certain psychological state" (p.135), and secondly that, "the meaning of a term (in the sense of "intension[7]") determines its extension[8] (in the sense that sameness of intension entail sameness of extension)" (p.136). He challenged these two assumptions, and proposed that meaning should be defined by specifying a normal form (or type of normal form) for the description. By this he means that meaning is partly socially determined, and that meaning is based on experience. The components of the normal form description of meaning are: syntactic markers, semantic markers, a description of the additional features of the stereotype and a description of the extension.

Putnam's (1975) instantiation of the normal form description for water is shown at Table 1.

Syntactic Markers	Semantic Markers	Stereotype	Extension
mass noun, concrete	natural kind, liquid	colourless, transparent, tasteless, thirst-quenching	H_2O (give or take impurities)

Table 1: Normal form description for water (Putman, 1975, p.191).

[7] Intension is the "concept" associated with the term. The timeworn example, of 'creature with a heart' and 'creature with a kidney', implies that the two terms have different intension, but the same extension (Putnam, 1975).

[8] Extension is the set of things the term is true of. For example, the extension of 'dog' is precisely the set of dogs. Putman (1975) argues that there are problems with this interpretation of meaning.

He comments, "If we know what a "normal form description" of the meaning of a word should be, then, as far as I am concerned, we know what meaning is in any scientifically interesting sense" (p.190). Lakoff (1987) said this approach was deficient for two reasons: firstly, only the propositional structure was included and the imaginative structure (metonymic, metaphoric and image schematic) were excluded, secondly, he only used a single representation for each category. Moreover, "mental representations for categories cannot be given meaning via their relationship to categories in the world" (*ibid.* p.372). Despite these deficiencies, Lakoff (1987) praised Putnam's (1975) early use of cognitive models, but asserted that, a cognitive model should also be able to account for the categorisation of phenomena in general.

Lakoff (1987) also discusses the fundamental problem of computational approaches to the study of the mind. However, the problems he describes of AI are somewhat similar to problems associated with determining semantic integrity. For example, often the need for meaningfulness is lost on the data modeller. The reason may be that when developing a data model, the data modeller does so with her/his own understanding of what the entities and business rules are supposed to mean. For this reason, the model doesn't seem meaningless to the data modeller because the entities and business rules are chosen with an *intended* interpretation. However, the model may not incorporate an adequate nonobjectivist account of what makes the data model meaningful to the person(s) whose thinking is being modelled. According to Lakoff (1987) an adequate theory must take into account how the *content* of a concept is related to bodily experience.

If the activity of data modelling is 'knowledge creating' then it involves an organisation of knowledge by means of structures which Lakoff (1987) calls idealised cognitive models, or ICMs. Each ICM is described as a gestalt, comprising four kinds of structuring principles: a propositional structure, an image-schematic structure, metaphoric mappings and metonymic mappings. He describes a simple ICM using the English word *Tuesday*. He suggests that *Tuesday* can be defined according to an idealised model that includes "the week is a whole with seven parts organized in a linear sequence; each part is called a *day*, and the third is *Tuesday*" (*ibid.* p.68). Another example provided is the concept of the weekend, which he says, "requires a notion of a *work week* of five days followed by a break of two days, superimposed on the seven-day calendar" (*ibid.* p.69). He says that this model of the week is idealised, as seven-day weeks do not exist objectively in nature, but are created by people.

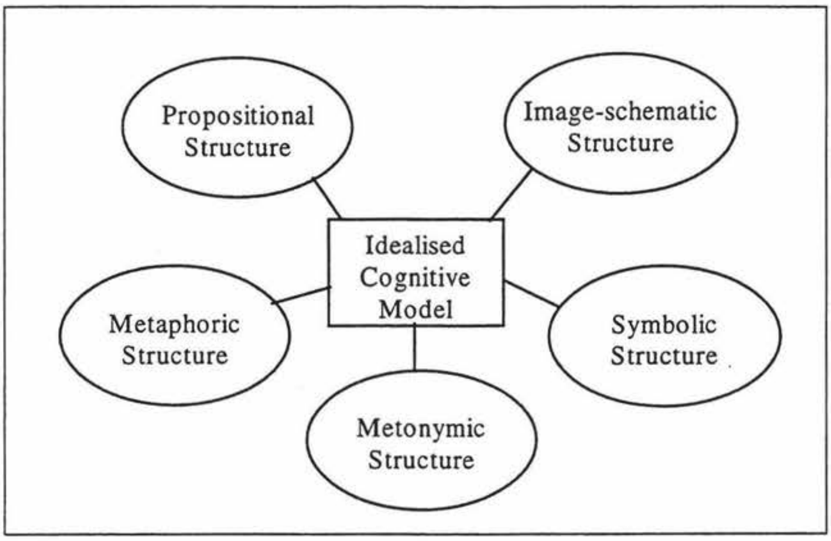


Figure 3: Structuring principles of ICMs
(Based on: Lakoff, 1987)

Figure 3 represents the five structuring principles of ICMs. The first structuring principle (or model[9]) is the propositional model; this model specifies the elements, their properties, and the associations amongst them. The example provided by Lakoff (1987) is a propositional model characterising knowledge about fire, which would include the fact that fire is dangerous. Whereas the second type of structure, image-schematic, specifies, "schematic images, such as trajectories or long, thin shapes or containers" (*ibid.*p.114). Metaphoric models are "mappings from a propositional or image-schematic model in one domain to a corresponding structure in another domain" p.114). Metonymy is the capacity to let one thing stand for another for some purpose, so a metonymic model can be either: social stereotypes, typical examples, ideal cases, paragons, generators, submodels or salient examples, that, according to Lakoff (1987) all produce prototype effects[10] of some kind. Like Putman's (1975) description of a stereotype as an idealised mental representation of a normal case, Lakoff (1987) defines social stereotypes[11] as something that can be

[9] Lakoff (1987) uses the terms interchangeably.

[10] A prototype effect is where certain members of a category are judged more representative of the category than other members (Lakoff, 1987). For example, "robins are judged to be more representative of the category BIRD than are chickens, penguins and ostriches" (*ibid.* p.41).

[11] An example of a social stereotype provided by Lakoff (1987) was Robins and Sparrows are typical birds (p.86).

used to stand for a category as a whole, but sometimes these are recognised as not being accurate. He remarks that these social stereotypes are "categories that function as stereotypes for other categories. An understanding of such categories requires an understanding of their role as stereotypes" (*ibid.*p.86).

Conceptual ICMs are characterised by Lakoff (1987) "independently of words and morphemes of particular languages" (p.289), however when linguistic elements are associated with conceptual elements in ICMs, a symbolic model results. He concludes this discussion by writing,

> "linguistic expressions get their meanings via (a) being associated directly with ICMs and (b) having the elements of the ICMs either be directly understood in terms of preconceptual structures in experience, or indirectly understood in terms of directly understood concepts plus structural relations" (*ibid.*p.291).

These ideas may be useful for analysing how a person interprets sentences derived from the design data model (refer to chapter seven for general strategies). Also, these structures are an attempt to answer questions such as: what is it about the human mind that allows it to categorise concepts in a particular way? Is there some general cognitive apparatus used by the mind that gives rise to categorisations of this sort? (Lakoff, 1987, p.113). These questions are relevant to data modelling because the activity of data modelling is a categorisation activity. According to this approach categories can be characterised using cognitive models of five types: propositional models, image-schematic models, metaphoric models, metonymic models and symbolic models.

If data modelling reflects some aspect of human reasoning, a corresponding cognitive semantic framework can be constructed using image schemas, metaphors and metonymies. Lakoff (1987) demonstrates cognitive semantics that covers the subject matter of predicate calculus, but then comments, "The resulting logic would apply to any subject matter that can be understood in terms of these schemas... it would provide an intuitively meaningful semantics" (pp. 366-367). Producing 'meaning' from a data model may be achieved by defining the components of the model according to Lakoff's (1987) schematic structures. Furthermore, these structures are similar to the semantic concepts described by Date (2000). Date (2000) claimed that semantic modelling includes, the identification of semantic concepts, the set of formal or symbolic objects, general integrity rules and finally the formal operators for manipulating the formal objects. The identification of semantic concepts[12] corresponds to the CONTAINER, PART-WHOLE, IDENTIFICATION and LINK schemas. However, the SOURCE-PATH-GOAL schema recognises that complex events have initial states (source), a sequence of intermediate stages (path), and a final state (destination), this relates to the intension of the data model. Table 2 provides an initial[13] description of how the semantics of a data model may involve metaphorical mappings based on some of the image schemas described by Lakoff (1987).

[12] The notion of semantic concepts as described by Date (2000) is relevant to this research, however an analysis of the other three steps is beyond the scope of this research.

[13] Further work is required to develop these ideas further. The development of image schemas for data modelling may be a fruitful area of future research.

Schema (Lakoff (1987))	Metaphorical Mapping	Structural Elements	Description
CONTAINER	Class. (the world is made up of entities).	INTERIOR, BOUNDARY, EXTERIOR	Each entity[14] is represented structurally by a CONTAINER schema.
PART-WHOLE	Sub class. Type versus instance.	A WHOLE, PARTS and a CONFIGURATION	All entities will have certain properties in common.
IDENTIFICATION	Unique identifier.	IDENTITY	Every entity has a special property that serves to identify that entity.
LINK	An entity can be related to other entity by means of relationships.	Two entities, A and B, and LINK connecting them.	To secure the location of two things relative to one another. Represents how the relational structure is understood.
SOURCE-PATH-GOAL	Intension. (Purposes are understood in terms of destinations, and purpose is understood as passing along a path from a starting point to an endpoint).	A SOURCE, a DESTINATION, a PATH and a DIRECTION.	Complex events have initial states (source), a sequence of intermediate stages (path), and a final state (destination).

Table 2: Image Schemas for a Data Model

Lakoff (1987) says that each schema has an internal structure and are understood in terms of direct experience. However, other more interesting forms of metonymic reasoning relevant to data modelling may be: social stereotypes, typical case, ideals, paragons, generators and salient examples. He remarks that these are normal activities involving the use of human reason, which involve "imaginative projections based on understanding an entire category in terms of some subpart of that category" (p.367). He continues,

> "In cognitive semantics, the study of the general forms of metaphoric, metonymic and image-schematic reason are no longer off-limits. This is human reason not transcendental reason. It can in principle be characterized with appropriate precision. It can apply to any subject matter that we can understand using image schemas, metaphors, and metonymies" (*ibid.* p.367)

ENTAILMENT is also important when describing meaning, because it is defined in terms of truth, which is in turn described in terms of understanding. If the data model is represented using NIAM or NaLER sentences (Atkins, 2000) a sentence is

[14] The use of the term entity here is not meant in terms of a relation in data modelling.

true if "it is true by virtue of what it means and how it is understood. Truth depends on meaningfulness" (Lakoff, p.294)

If data modelling is a form of categorisation, then it is a matter of both experience and imagination, as Lakoff (1987) writes,

> "human categorization is essentially a matter of both human experience and imagination - of perception, motor activity, and culture on one hand, and of metaphor, metonymy, and mental imagery on the other" (p.8). Consequently, he says, "human reason crucially depends on the same factors, and therefore cannot be characterized merely in terms of the manipulation of abstract symbols" (*ibid.*p.8).

In addition, Lakoff (1987) describes the notion of concepts, for example the concept flight attendant can be characterised relative to an airline scenario (because for every concept there can be a corresponding category). A category is "those entities in a given domain of discourse that the concept (as described by the cognitive model) fits" (*ibid.*p.286). Concepts characterised in a cognitive model using necessary and sufficient conditions generate classically defined categories, whereas prototype effects can arise in a number of ways such as: metonymy, radial categories, generative categories and graded categories (refer Lakoff, (1987), pp.286-289). These ideas may be useful for determining semantic integrity, in particular for understanding the meaning or the semantic content of the data model during data warehouse design. Using the theory of categorisation and the notion of prototype effects for achieving semantic integrity may be an interesting and fruitful area for future research. Despite this, the theory of categorisation "makes predictions about what human category systems can and cannot be like. It does not predict exactly what will be in a given category in a given culture or language" (*ibid.* p.96).

Levels of Meaning

Lakoff (1987) asserts that meaning is based on human perception, interaction and understanding. This perception, interaction and understanding is also captured in the framework proposed by Mingers (1995), which describes different levels of meaning for both the generation of meaning from signs, and for the production of signs from meaning.

Generation of meaning from signs

The first level of meaning refers to how a person (receiver) comes to **understand** the primary meaning of a sign or linguistic message. This is

> "the level of understanding that can be expected from all competent speakers of a language - all those who share a particular language or symbol system. It corresponds to the semantic content... that is, the digitialized information without its analogue nesting" (Mingers, 1995, p.299).

He comments that the main validity claim involved with this meaning is comprehensibility. However, this level of meaning is not always easy to obtain, because it can be difficult to understand a particular utterance[15].

> "Often a negotiation or interchange is necessary to establish it, especially when speakers are not straightforward or sincere, and employ irony or sarcasm to negate the surface meaning. If comprehensibility is a problem, it may reflect a lack of adequate *structural coupling* (Maturana, 1978) between speaker and receiver - the signs do not have common connotations - or it may call into question other validity claims, particularly sincerity" (*ibid.*p299).

Likewise, Habermas (1984) remarks that to reach an understanding requires that,

> "[A]t least two speaking and acting subjects understand a linguistic expression in the same way. The meaning of an elementary expression consists in the contribution that it makes to the meaning of an acceptable speech act. And to understand what a speaker wants to say with such an act, the hearer has to know the conditions under which it can be accepted" (p.307).

He claims that an agreement of understanding is achieved between hearer and speaker through the communicative intent of the speaker. In order to come to an understanding with a hearer about something and thereby to make her/himself understandable, the intent of the speech act must be: *right* (with respect to the given normative context), *true* (the hearer will accept and share the knowledge of the speaker) and *truthful* (the expression of beliefs, intentions, feelings and desires). Therefore, according to Habermas (1984) when a person makes a statement, asserts, narrates, explains, represents, predicts or discusses something, s/he is looking for an agreement with a hearer based on the recognition of a truth claim. He classified the speech acts according to cases of communicative action: conversation, normatively regulated action and dramaturgical action.

The second level of meaning '**connotation**' incorporates beliefs and implications that the person (receiver) associates with the primary meaning (Mingers, 1995). This is where a person's knowledge and experience influence the meaning derived. In the context of linguistics Putman (1975) addressed this as the division of linguistic labour. He writes,

> "Every linguistic community exemplifies the sort of division of linguistic labor just described, that is, possesses at least some terms whose associated "criteria" are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets" (*ibid.* p.146)

Mingers (1995) also remarks that connotation "is not primarily individual but will be differentiated between groups of people" (p.300). For example, the different stakeholders such as data producers, data custodians, data consumers and data managers (Giannoccaro *et al.* 1999), all form differentiated groups who share experiences that are unavailable to outsiders (Mingers, 1995, p.300). This second level of meaning is concerned with Habermas's (1984) validity claims of truth and rightness. In line with these validity claims, the following questions are suggested: is the propositional content of the sign actually correct? Does the state of affairs actually exist? Are its claims about social rules and roles acceptable?

The third level of meaning '**intention**' is both the individual meaning for a particular person plus the implications for action. This meaning level examines what intention it will lead an individual to have. For example personal experiences, feelings and motivations will determine resulting activity, which may only be remembering for future use (*ibid.* p.300). The validity claim applicable for 'intention' is sincerity, this can be measured by determining; whether the source is truthful, whether the speaker meant what they said and by determining the reliability of the speaker.

Production of signs from Meaning

The obverse of meaning generation is the production of signs from meaning, which involves creating an utterance or gesture (an analogue sign) from a digital meaning. The three levels of meaning associated with this process are: intention, generation and action.

[15] An utterance is some combination of signs or symbols produced at a particular time with some *intent.* This could range from speech or writing, a gesture or input into an information system (Mingers, 1995, p.292).

The first stage relates to the **intention** of the sender or producer of the sign, as mentioned before sincerity is the primary validity claim. The second level '**generation**' involves the conversion of an intention into a specific form that can be represented by signs or utterances. Mingers (1995) says there are a wide range of possibilities here, where the transition from intention to action is almost instinctive (e.g. the reaction to fear) or a slow process that is highly political. He remarks that, generation "always occurs within the context of particular forms of life and draws on social structures and constraints such as language, practices, skills, resources and power" (*ibid.* p.301). Generation is concerned with Habermas' (1984) validity claims of rightness, truth and effectiveness.

The third meaning '**action**' is where a comprehensible utterance or sign must be generated. Mingers (1995) remarks that competence in the semantic and syntactic rules of the language or sign system is necessary if the sign is to be understood. This is what he describes as the production of an analogue sign from a digital meaning.

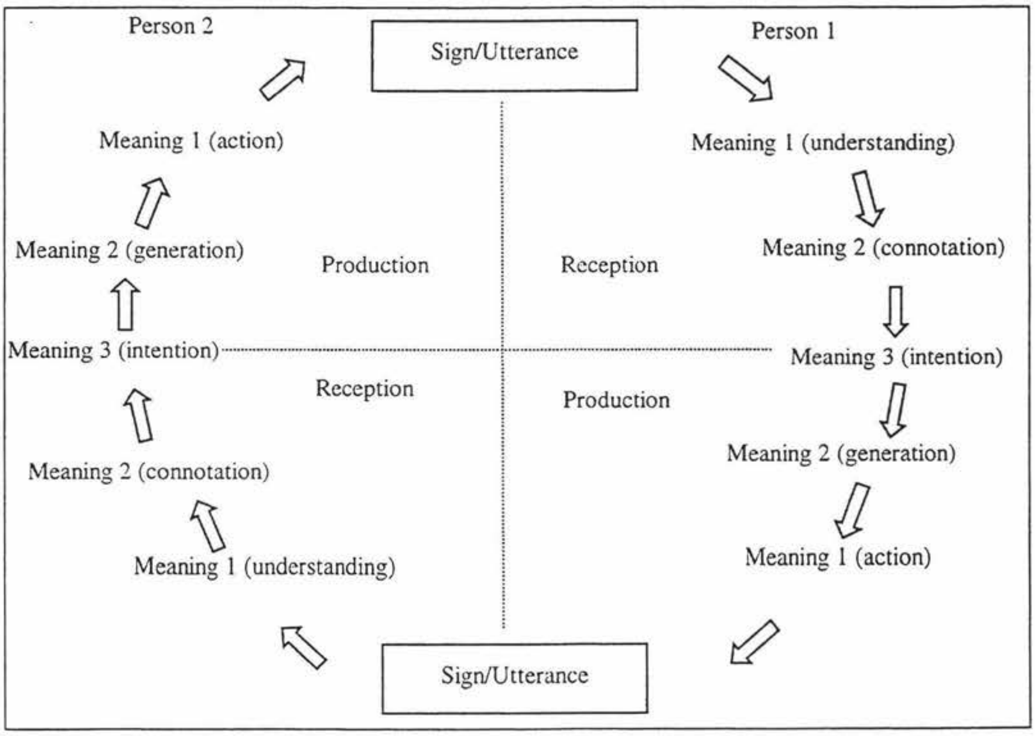


Figure 4: Levels of meaning (Mingers, 1995, p.299)

Although the levels of meaning described are tightly interwoven, the framework is useful for analysing both the subjective and intersubjective dimensions of meaning (illustrated in Figure 4). The framework is useful as a mechanism for revealing the different understanding, connotation and intentions the data consumers may have in the context of data warehousing. Likewise, as a mechanism for revealing the intention, generation and actions of the data producers. This is important because to make sense of the information in the data warehouse "a lot of contextual knowledge is needed; usually this knowledge is not stored within the computer system. Instead system designers implicitly rely on culturally shared and accumulated stocks of knowledge" (Tuomi, 1999, p.110).

Table 3 describes the meaning levels from the viewpoint of a receiver gaining meaning from signs. In the context of this research this may be the viewpoints of the end users' gaining meaning from the information in the data warehouse. Mingers (1995) says it is just as important to examine the producers of signs and utterances, and the relations between their meanings and the signs produced. Table 3 also describes the meaning levels from the viewpoints of the producers or senders of the sign. In the context of this research this may be the viewpoints of the data warehouse designers.

Meaning	**Description**	**Validity Claim (Habermas, 1984)**	**Problem**
1. Understanding (generating meaning from signs)	Comprehending the data. (understanding the semantic content) dependant on the data consumer's stock of tacit knowledge.	Comprehensibility, Truthfulness (sincerity)	Data difficult to understand, out of context. Lack of adequate structural coupling (Matuarana, 1978)) between speaker and receiver.
2. Connotation	Relating initial understanding with other knowledge and experience. (intersubjective)	Truth Rightness (legitimacy)	Assumptions made are inappropriate. Conventions used are not standard. (Not a conscious process that can be measured).
3. Intention	The individual meaning for a particular person and the implications of that meaning for action. (subjective)	Sincerity	May lead to incorrect (or unintended) decisions or actions.
1. Intention (producing signs from meaning)	The intention of the sender or producer of the sign.	Truthfulness (sincerity)	The intention of the producers may not match the 'need' of the receivers.
2. Generation	Converting an intention into a specific form.	Truth Rightness (legitimacy) Effectiveness (in the case of strategic action).	May be constrained by language, practices, skills, resources and power. Highly political.
3. Action	A comprehensible utterance or sign is produced.	Comprehensibility	Lack of competence in understanding the semantic and syntactic rules of the language or sign system.

Table 3: Levels of Meaning

Strategies for achieving semantic integrity are discussed in chapter seven in the context of data modelling for the data warehouse (based on the ideas presented in Table 3).

Intersubjective Meaning in Data Modelling

The research of Mingers (1995) provides the groundwork for constructing a framework to assist with abstracting meaning **out** of the data model. As Marche (1993) discovered through his work on measuring schema stability of data models, "people are so effective at ...projecting meaning onto the data structures rather than abstracting meaning out of them" (p.45). In addition, Batra and Srinivasan (1992) comment that "an important concern for the researcher is how the modelling process may be studied in its entirety, i.e. representation coupled with manipulation" (p. 409).

The framework at Figure 5 describes intersubjective meaning in data modelling (in this context a data model[16] is similar to a number of symbolic utterances). Moreover, a data model by itself is not information, but it may *carry* information about a particular set of data. It is an abstraction device "that allows us to see the forest (information content of the data) as opposed to the trees (individual values of data)" (Tsichritzis & Lochovsky, 1982, p.5).

[16] As opposed to the term Data Model (capitalised), which generally refers to the "set of conventions used to represent a simplified, formal and highly abstracted view of data" (Atkins, 2000, p.30).

Figure 5 also portrays how meaning may be generated from a model, through understanding, connotation and intention. In addition, how either a subsequent model or knowledge may be generated from meaning through intention, generation and action.

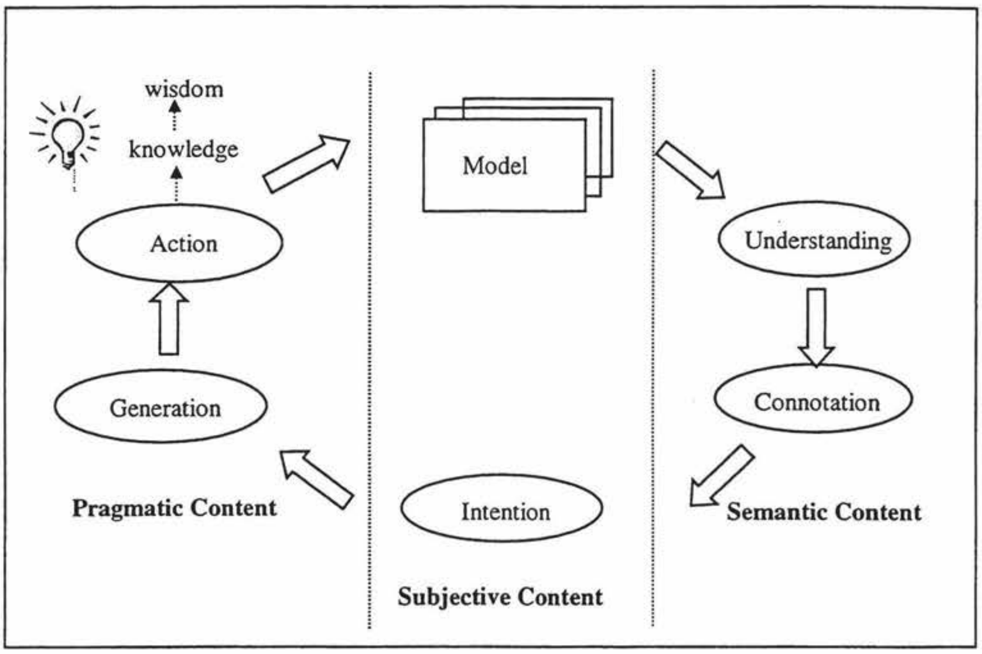


Figure 5: Intersubjective Meaning in Data Modelling

The implication of this framework is that we can only ever expect a user to reach the points of action, knowledge or wisdom if they have sufficient understanding of the semantic content of the data model. Achieving this understanding may be problematic as Artz (1997) highlights, "people do not possess an innate mental logic for understanding data dependencies and their implications" (p.30). Defined in this context 'understanding' is a prerequisite for action. However, this research explores the importance of semantic integrity in data warehouse design, this theory, based on semiotics is useful for clarifying the problem.

Generation of meaning from a data model

1. Understanding: Like the classical approach to categorisation, the generation of meaning from a data model has been overlooked because the purpose is most often been to categorise entities in terms of the shared properties of the entity type, and has omitted the peculiarities of human understanding.

To understand a data model the data consumer must realise both the importance of the activity and the importance of their participation. Problems may also arise such as conflicting interpretations amongst the data consumer community. Factors such as: general experience, prior in-house experience (of existing systems or databases), length employed, job role, cognitive style and confidence all impact on the degree of understanding a data consumer may have.

The majority of data consumers are only interested in understanding something that is applicable to their domain or job role. Darke and Shanks (1997) note that "each stakeholder group often understands and accepts responsibility only for the part of the application domain relevant to its specific interests" which can be problematic when the "differing views and perceptions of the problem situation" are integrated "into a coherent set of requirements" (p.214). Likewise, Benyon (1997) writes, "The same signs may be interpreted by different people in different ways depending on their existing knowledge and their ability to interpret the signs" (p.4). Therefore, to achieve semantic integrity necessitates the existence of an overlapping meaning structure between the data modellers (or data warehouse designers) and data consumers (end users). The ability to achieve this is to some degree dependent on the depth of knowledge the end user has of the domain and their ability to communicate that knowledge.

Lakoff (1987) says that image schemas and metaphorical models are required to represent the meanings of expressions. The meanings of the concepts in a data model can also be represented using image schemas and metaphorical models (refer Table 2).

Assuming the conceptual and logical modelling for the source database(s) has already taken place, the data warehouse designers are faced with the difficulty of verifying these original designs and tailoring them for data warehouse design. To reach an understanding of the resulting data model it is not necessary to expose the data consumers to the formality of an E-R model. On the contrary, research has been undertaken that suggests this approach may be problematic (Atkins, 2000). If a data model is viewed as a vehicle for communicating *meaning* through understanding, connotation and intention, it is important to use a language that is familiar and easy to use. A useful technique to facilitate communication in the design phase[17] is the natural language technique 'NaLER' proposed by Atkins and Patrick (2000). Additionally this technique is extremely fruitful for validating the perceived semantic content of the model.

2. Connotation: This is where the data consumer will relate the initial understanding described above to other knowledge and experience. Mingers (1995) says this is not a wholly conscious process, therefore, it is difficult to determine strategies for enhancing this meaning level. However, the data model should be validated with the participation of data consumers. This can be achieved partly by questioning the presumed validity claims such as verifying whether: the assumptions made are valid, the business rules modelled are acceptable and correct, the espoused motives of the end users'/data modellers' are sincere.

3. Intention: It is important to consider the intention of the data consumers, documenting this at the data modelling stage is essential as it may uncover complexities (exceptions) missing in the model. Likewise, the data modeller should try to determine the actions it will lead the end users to have and this intention should be supported through the data model. For example, asking questions like the following may help determine intention: what do they intend to do with the information? What type of reports do they require? What level of detail is needed? The end users should ask questions such as: will the information carried by the data model be truthful? Will the information in the data warehouse be reliable based on the business rules in the data model? Will the data model provide information that can support all business requirements?

Producing a data model from meaning

For the purposes of this research it is assumed that some form of conceptual data modelling has occurred during the analysis phase of the project, prior to the data warehouse design. Hopefully too the conceptual modelling activity resulted in a number of data models that represent the different stakeholder interpretations of the

[17] Although Atkins (2000) proposes the use of NaLER for developing a conceptual data model, it may also be a useful technique for verifying any data model. Although the dimensional model has been alluded to, this discussion has not specifically, addressed the data model(s) type e.g. conceptual or design, this is largely due to the range of data model types used by practitioners for data warehouse design (e.g. ER, E-R/Relational Hybrid, Dimensional, or Relational).

business domain. Also, although it is outside the scope of this project to investigate the role of conceptual data modelling, previous research suggests (Atkins, 2000) that the use of NIAM and the development of an object-role-model is the only way to usefully validate the analysis model, apart from building the actual database. An important area of future research could examine the impact of this notion and perhaps investigate whether the only 'real' way of verifying semantic integrity of a data warehouse is to trace back to the conceptual 'utterance' and validate this through the use of a natural language approach such as NIAM.

Nevertheless, it is still important to analyse how a design data model is produced from an existing meaning structure. The epistemological dimension that Hirschheim *et al.* (1995) referred to as "how developers inquire into object systems and see phenomena in them" (p.21) can be equated with the problem of producing a data model from meaning. Using Mingers' (1995) meaning levels of intention, generation and action, the problem can be analysed in detail.

1. Intention: Because meaning is based on experience (Lakoff, 1987; Putnam, 1975), the intention of the data modeller is based on the understanding of what has been indirectly communicated to the data modeller by the data consumers experience. Therefore the problem is that 'meaningfulness' to the data modeller is very indirectly based on the experience of others. The intention includes determining for example what is to be modelled and why?

According to Lakoff (1987) conceptual categorisation choices are "at least in part, determined by the bodily nature of the people doing the categorizing, rather than solely by the properties of the category members" (p.371). Also the conceptual categories "have properties that are a result of imaginative processes (metaphor, metonymy, mental imagery) that do not mirror nature" (*ibid.*p.371).

Using **metonymy** or reference-point reasoning a data modeller may determine intention. A metonymic model has the following characteristics:

- There is a "target" concept A to be understood for some purpose in some context.
- There is a conceptual structure containing both A and another concept B.
- B is either part of A or closely associated with it in that conceptual structure. A choice of B will uniquely determine A, within that conceptual structure.

- Compared to A, B is either easier to understand, easier to remember, easier to recognise, or more immediately useful for the given purpose in the given context.
- A metonymic model is a model of how A and B are related in a conceptual structure; the relationship is specified by a function from B to A. (Lakoff, 1987, pp.84-85).

These characteristics infer that B may be used to stand, metonymically, for A. Types of metonymy are; social stereotypes, typical cases, ideals, paragons, generators, and salient models. These may be useful classification techniques for supplementing the requirements definition phase of systems analysis. The data modeller may use metonymic models to aid understanding of the problem domain.

A **social stereotype** is when a judgement is made about people or a situation. Social stereotypes (according to the stakeholders perception) may be used to describe parts of the data model. However, Lakoff (1987) warns, "they are usually recognised as not being accurate, and their use in reasoning may be overtly challenged" (p.85). Therefore, recognising when a person is using a social stereotype to describe a category as a whole may help avoid placing too much importance on such descriptions.

A **typical category** is when inferences from typical to atypical cases are made, based on knowledge of the typical. It is very normal to categorise things in terms of typical cases, however, a 'good' data modeller would try to model typical cases as well as those cases that were atypical.

A category can be understood in terms of abstract **ideal cases** which involve making judgements of quality and planning for the future, for example, examining the current data quality problems, and looking at what an ideal situation might be. According to Lakoff (1987) a lot of cultural knowledge is organised in terms of ideals, for example ideal jobs, ideal workers and ideal bosses.

Paragons include making comparisons and then using that comparison as a model for behaviour. **Generators** define concepts by principles of extension, where the members of a category are defined or generated by the main members. **Salient examples** include using familiar, memorable examples to understand categories. The data modeller may use many types of examples and not only salient examples to understand the categories. Atkins (2000) suggests instantiating NaLER sentences with examples to verify the model,

> "example sentences are generated from the fact-types and instantiated by relevant valid examples taken from the UoD. Where NaLER is being used within a system development process, the examples should be taken from the analysis documentation wherever possible, rather than generated in isolation by the designer. Any examples that are created by the designer should be verified by the user" (p.157).

2. *Generation:* This meaning level is concerned with conversion of intention into specific action. The main validity claims are rightness, truth and effectiveness. This meaning level may be associated with the data modelling activity. However, the activity of data modelling for the data warehouse is described in chapter three of this report.

3. *Action:* A comprehensible data model is produced for the data warehouse. This implies competence in the semantic and syntactic rules of the data model if the data model is to be understood. Action addresses the representation question - how is the result presented? A detailed description of this meaning level is outside the scope of this research, however, information was collected regarding the 'action' resulting from the use of the data warehouse (refer to chapter seven).

Data model quality: The wider context

The fundamental reason for examining meaning in detail has been to address the need for quality in data modelling. Although this research is only looking at one aspect of data quality, that of semantic data quality, it is useful to examine previous research in data quality. For example, Shanks & Darke (1998a) have also used semiotic theory to develop a framework for understanding data quality in a data warehouse. However, their framework covers a much broader scope where the key component is a set of data quality goals for the four semiotic levels, syntactic, semantic, pragmatic and social. Their framework uses semiotic theory to examine quality for both the **content** and the **structure** (metadata) of a data warehouse. When they are referring to the data warehouse structure they are referring to the particular notation used, or the "set of statements in some language" (*ibid.* p.124), for example the star schema notation. The data warehouse content is described as "the symbolic representation of values of the elements defined in the data warehouse structure" (*ibid.* p.124). The framework components are illustrated at Figure 6.

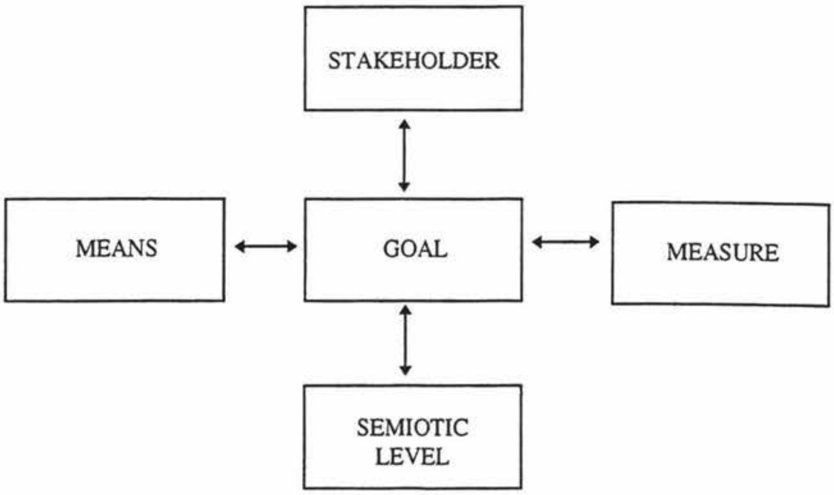


Figure 6: Components for Understanding Data and Metadata Quality in a Data Warehouse (Shanks and Darke, 1998a, p.125)

For each of the semiotic levels, syntactic, semantic, pragmatic and social they provide a goal, a means to achieve the goal, and how to measure the goals. However, for the purposes of this research it is only necessary to discuss their findings for the semantic semiotic level.

As mentioned earlier, the goals for semantic quality are **completeness** and **validity**, in the context **structural** quality. According to the goal of completeness, "each stakeholder may have a different view of how complete the data warehouse structure is" (Shanks and Darke, 1998a, p.125). Moreover, "a valid data warehouse structure does not include any elements that are not in the stakeholder's conceptualisations of the domain of interest" (*ibid.* p.125). The means to achieve completeness are: correctness, stakeholder training, stakeholder participation and prototyping. The measures they suggest for completeness and validity were extracted from Moody and Shanks (1998), these were: expert rating of the data warehouse structure, and the comparison of the data warehouse structure with generic models.

Shanks and Darke (1998a) also applied their framework to the actual **content** (or data) by defining the goals, means and measures for each of the four semiotic levels. They made use of the research of Kahn *et al.,* (1997) and Tayi and Ballou (1998) to define the goals for semantic quality of **completeness** and **accuracy**. They define completeness in this context as, "the degree to which the data warehouse *content*[18]

[18] Emphasis added.

represents each of the data values in the stakeholder's conceptualisations of the domain of interest. Each stakeholder may therefore have a different view of how complete the data warehouse content is" (Shanks and Darke, 1998a, p.126). They also define accuracy as the "degree to which values of data elements in the data warehouse map on to the domain of interest" (*ibid.* p.126). Their suggestions for achieving a complete and accurate data warehouse are: consistency, stakeholder training in both the data modelling language and the domain of interest, they also recommend, "reducing the number of transformations and transcriptions of data from when it is first captured to when it is stored in the data warehouse" (*ibid.* p.126). To measure completeness and accuracy techniques such as population sampling and comparison are recommended.

The framework of Shanks and Darke (1998a) is very useful because they explain both the goals to aim for within the particular semiotic level, and a means for measuring the goal. They suggest that the goal for **pragmatic** quality is **understanding**, after Lindland *et al.,* (1994), whereas Mingers (1995) describes understanding as the "first level of meaning in which a receiver comes to understand the *primary* meaning of a sign or linguistic message" (p.299). Mingers (1995) claims that this level of meaning corresponds to the **semantic** content[19], and that the main validity claim is **comprehensibility**, this may be due to the influence of Dretske (1981) who asserted that, instead of meaning generating information, information is seen to generate meaning. Consequently, understanding corresponds to the semantic content of a sign or linguistic message in Mingers (1995) framework. (However, it is apparent there is some overlap between the goals for semantic data quality and the goals for pragmatic data quality).

Two additional goals for the semantic semiotic level might be: **meaningfulness** and **comprehensibility**. These goals may be appropriate because 'semantics' deals with the issue of meaning, or the relationship between signs and what they are supposed to represent (Stamper, 1987). These goals are necessary to capture the importance of intersubjective meaning and relate directly to understanding the semantic content of a sign or linguistic message. As described in chapter one 'comprehensibility' may be appropriate in terms of both the structure and the content, however, 'meaningfulness'

[19] This is the "digitalized information without its analogue nesting" (Mingers, 1995, p.299)

may be appropriate in terms of the content of the data warehouse. Although Shanks and Darke (1998a) include 'understanding' as the goal for the pragmatic semiotic level it is important to also recognise that this goal relates to the goals of 'meaningfulness' and 'comprehensibility'. The means to achieve these goals are described in chapter eight.

Atkins (2000) also addressed the need for quality in the context of conceptual data modelling. She developed a framework based on the work of Krogstie *et al.* (1995) and Moody and Shanks (1994). The framework was utilised in her research for assessing the final products of the INTECoM development. Her quality framework is shown at Figure 7.

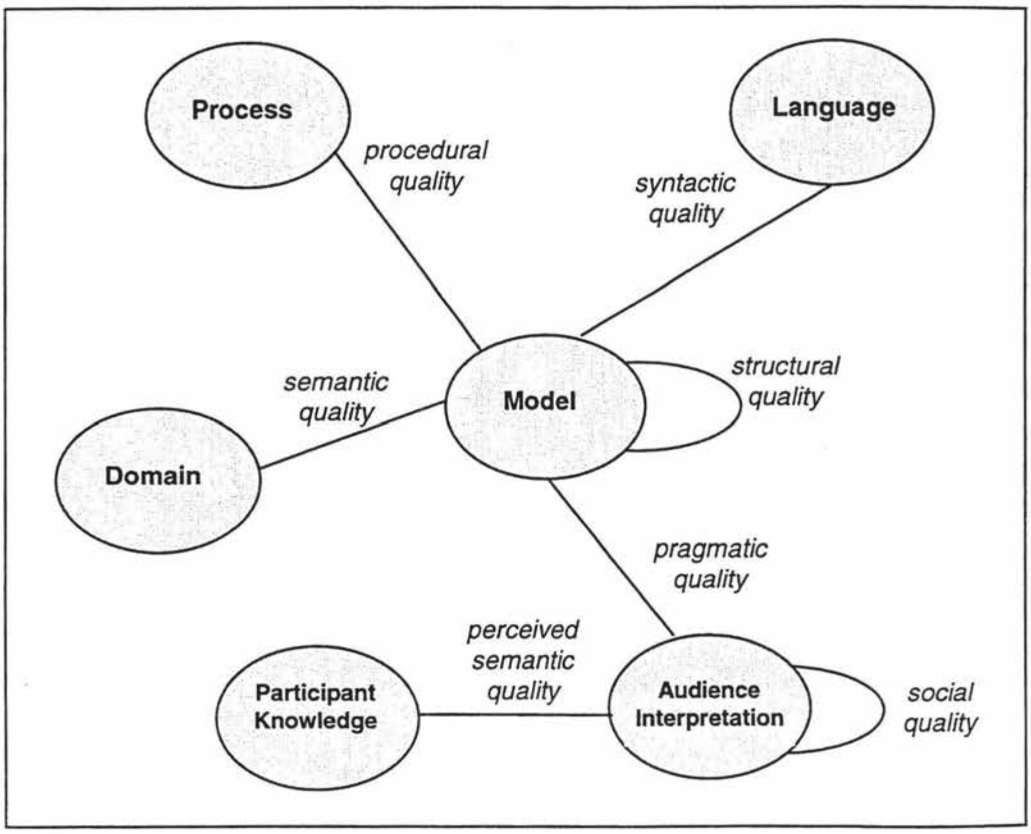


Figure 7: INTECoM - Quality Framework (Atkins, 2000)

She stipulates that semantic verification is a pertinent step when creating the analysis model using the INTECoM framework. However, as she demonstrates in the quality framework, this can only ever be a verification of the *perceived* semantic quality.

> "The user is required to provide verification of the semantic content of the model, e.g. that the fact types and example sentences recorded in the model are meaningful and accurate within the domain appropriate to that user" (*ibid.* p.138).

Tozer (1999) argues the importance of data quality in the data warehouse through a discussion of metadata management. He claims that metadata, which defines and controls the behaviour of data in the data warehouse, should be kept under control and consistent. He recognises long term problems when managing metadata, in particular, "how to ensure that the behaviour of data - which arrives within the warehouse environment and is transformed and loaded into separate data marts, restructured by OLAP ad DSS tools, and then separately interpreted by users - remains consistent" (*ibid:* p131). Commendably, he notes the possibility for individual interpretations of the business rules,

> "the possibility exists for reinterpretation of the business rules that govern this behaviour, introducing inconsistencies. These inconsistencies will be carried forward into the reports, leading to errors of judgement, misreporting of statistics, and time wasted in trying to reconcile apparent conflicts in figures" (*ibid*, p131).

Tozer (1999) recommends one of four approaches to ensure metadata consistency. While he makes some useful suggestions, the approaches he suggests are based on the tools used to manage the metadata.

Summary

Semantic data quality is crucial in data warehousing, because the information carried by a data warehouse is often expected to convey the same, or at least similar, meanings to the various stakeholder groups. A framework based on the research of Mingers (1995) has been devised to describe intersubjective meaning in data modelling. The purpose is also to help achieve semantic integrity during data warehouse design. The framework illustrates this process by describing the levels of meaning relevant to: the generation of meaning from a data model and the production of a data model from meaning. Implicit in the framework is the assumption that a user can only ever reach the points of action, knowledge or wisdom if they have sufficient *understanding* of the *semantic content* of the data model. Therefore, this framework describes the stages involved for reaching these levels. Of primary relevance to this research is the generation of meaning from a data model, because designing a data warehouse should involve a validation activity and therefore an understanding of the source data model(s).

Furthermore, two existing quality frameworks have been analysed and utilised in this research (Shanks and Darke, 1998a; Atkins, 2000). Firstly, extending the work of

Shanks and Darke, (1998a) by the addition of two goals: 'meaningfulness' and 'comprehensibility' for the semantic semiotic level. This is to capture the importance of intersubjective meaning and to provide scope for understanding the semantic content of a data model. Secondly, the quality framework of Atkins (2000) has been used for describing the pilot case study findings.

The framework illustrated at Figure 5 will be used in this research as the conceptual structure for describing the case study findings and for presenting strategies for achieving semantic integrity through understanding, connotation and intention.

3 Data Modelling for the Data Warehouse

"Star schemas are really physical, not logical, though they are talked about as if they were logical. The problem is that there is really no concept of logical design, as distinct from physical design, in the star schema approach" (Date, 2000, p.714).

Introduction

The meaning stakeholders derive from the information represented by the design data model defines the semantic integrity of the data warehouse. Lakoff (1987) too says, "Meaning involves what is meaningful to us. Nothing is meaningful in itself" (p.292). Subsequently, the design data model cannot be, by itself, wholly semantically accurate.

In 1970 Codd proposed the concept of the data model, which Date (2000) defines as consisting of three components: a collection of object types, a collection of operators and a collection of general integrity rules. Specifically, he writes:

> "A **data model** is an abstract, self-contained, logical definition of the objects, operators, and so forth, that together constitute the *abstract machine* with which users interact. The objects allow us to model the *structure* of data. The operators allow us to model its *behavior"* (p.14).

Furthermore, he points out that, "current data models are not totally devoid of semantic features. For instance, domains, candidate keys, and foreign keys are all semantic features of the relational model as originally defined" (*ibid.* p.419). He also notes that the extended models developed to specifically address semantic issues are "only slightly more semantic than earlier models" (*ibid.* p.419). While these models may be termed 'semantic', they have not managed to "capture all of the semantics of the situation under consideration" (*ibid.* p.419).

Date (2000) contrasts the ideas of semantic modelling (using the E/R model) with the normalisation discipline and concludes that the overall design approach should include both top-down design (the E/R approach) and normalisation[20]. He quite rightly says, "the normalization discipline has absolutely nothing to say about how we arrive at those large relvars in the first place" (*ibid.* p.438). Moreover, Date (2000) recognises that semantic modelling is "not nearly so rigorous or clearcut as the further normalization discipline" (p.438) because, "database design is still very much a subjective exercise, not an objective one; there is comparatively little by way of really solid principles that can be brought to bear on the problem" (*ibid.* p.438). The ideas proposed in chapter two relate to the meaningfulness of the information provided by a data warehouse, which are in addition to the more formal, semantic, features of the relational model: domains, candidate keys, and foreign keys.

According to Date (2000), Peter Chen's (1976) entity/relationship (E/R) model is an application of a semantic model. The original E/R paper proposed both the E/R model and the E/R diagramming technique, yet Date (2000) claims that the popularity of the E/R model can be largely attributed to the diagramming technique. Also, he remarks that, "it is quite possible to use E/R diagrams as a basis for *any* design methodology" (*ibid.* p.438). For example, Atkins (2000) discusses the E-R/Relational hybrid model, which is also promoted by writers, such as Simsion (1994) and Benyon (1997). Atkins (2000) wrote of the hybrid model,

> "It uses E-R, or more commonly, extended E-R constructs (Cattell, 1994) and notation to provide a graphical representation of a normalised relational structure, is generally characterised by being in third normal form, having no relationship attributes and resolving all many-to-many relationships. Such a model is often termed an E-R model, or sometimes an E-R diagram, but clearly does not conform to Chen's (1976) original definition and it is this hybrid that should perhaps be the target of many of the criticisms currently levelled at the E-R Model" (p.45).

Throughout the data warehousing literature many authors debate the appropriateness of E/R modelling for data warehouse design (e.g. Devlin, 1997; Inmon, 1992; Kimball, 1996; Sterling, 1997). However, the E/R models referred to are in fact some form of relational model (E-R/Relational hybrid model) and not conceptual

[20] Date (2000) describes normalisation very loosely here as "reducing large relvars to smaller ones; it assumes that we have some small number of large relvars as input, and it manipulates that input to produce a large number of small relvars as output" (p.438)

E/R models. In fact, the authors are most often debating the inappropriateness of normalisation for data warehouse design.

The Data Warehouse

The term data warehouse originated in the late 1980s though the concept is older, for example, data warehousing projects were in existence in 1970 according to Hay (1997). Date (2000) succinctly says that a data warehouse is "a special kind of database" (p.709). He explains the motivation for a data warehouse as decision support data which,

> "needs to be collected from a variety of operational systems (often disparate systems) and kept in a data store of its own on a separate platform. That separate data store is a *data warehouse*" (*ibid.* p.709).

Similarly Shanks *et al.* (1997) state that a data warehouse is "a set of databases created to provide information to decision makers" (p.350). While Inmon (1993) defines a data warehouse as a subject-oriented, integrated, nonvolatile, time-variant collection of data organised to support management decisions.

While there is a proliferation of the practitioner literature on data warehousing, there is little formal academic research in data warehousing. From a practitioners point of view Devlin (1997) remarks, "Data warehousing provides the self-consistent and well-understood data needed to manage the business both as a whole and in its individual parts" (p.44). However, as the case study in this research shows, the data warehouse on its own does not provide 'self consistent' and 'well-understood data' (refer to chapter seven). Also, many confusing terms are used throughout the practitioner publications, for example architecting, data disciplines, entity relation data models and analytical database design (Devlin, 1997; Mattison, 1998; Kimball, 1998; Fong and Zeng, 1997). Adding to this confusion are the numerous terms used to describe the data modelling approach. For example, star schema, snowflake schema, constellation schema, blizzard (or snowstorm) schema, dimensional model, dimension map, hypercube design, and multidimensional model.

The most rigorous theoretical discussion on data warehousing can be found in Date (2000), where he devotes a chapter to decision support. In particular he discusses: data warehouses, data marts, dimensional schemas, star schemas, online analytical processing (OLAP) and multi dimensional databases. Most importantly for this research, he recognises that although most decision support databases are read-only, the designers should not overlook the semantic value of the constraints because,

> "the constraints serve to define the meaning of the tables and the meaning of the overall database. Declaring the constraints thus provides a means of telling users what the data means, thereby helping them in their task of formulating queries" (Date, 2000, p.699).

Data Modelling for the Data Warehouse

The three model architecture (conceptual model, logical model, internal or physical model) has been discussed in data modelling for some time. However, this was developed with transaction processing systems in mind. Since data warehouses are developed for analytical processing the concept of the dimensional model or star schema has been added.

Shanks *et al.* (1997) point out that the concept of data warehousing is not new, and indeed dimensional modelling is not new either. The dimensional model is based on the principles of relational data modelling (Codd, 1970). The main *fact* table is similar to an intersection entity (or resolution entity) in the relational model, because the primary key of the fact table is made up of the foreign keys from each dimension table. (However, this similarity does not to suggest that the approach adheres to good design practice as promoted by the relational model). Winsberg (1998) also notes the similarity when he says: "A star schema is a relational design with one fact table and many dimension tables" (p.519). Likewise, Date (2000) remarks, "a *simple* star schema... can look very similar (even identical) to a good relational design" (p.713). However, it must be emphasised that many star schemas are not equivalent to relationally correct, logical designs.

Winsberg (1998) describes three types of data warehouses that have emerged since the late 1980s: the enterprise data warehouse, the operational data store and the data mart. As the name implies, the enterprise wide data warehouse, provides a central data base for decision support throughout the enterprise. An operational data store is similar in scope to the enterprise data warehouse, but is updated by the system regularly, so it can operate like a transactional system. A data mart is a subset of the overall warehouse, with data specific to the needs of a user or a specific business unit. Therefore, the main difference between a data mart and an enterprise data warehouse is scope. However, a combination or 'multitier' warehouse approach can be adopted, which is where an enterprise warehouse and several data marts coexist together, this is similar to the notion of the corporate information factory (Inmon *et al.*, 1998).

Interestingly, Winsberg (1998) writes,

> "A star schema is essentially the same as a hypercube design for a multidimensional tool. The fact table is logically equivalent to a hypercube, except that the dimensions appear explicitly as values in columns. One difference is that a star schema may contain more dimensional information" (p.521).

Therefore, the main difference is between the tools that implement the models. Although, like most of the literature on data warehouse design these differences have little to do with the logical structure of the star schema and hypercube models.

Winsberg (1998) describes four separate components to the problem of modelling for the data warehouse. These components are: the dimensional model, the tabular model, source documentation and the enterprise model shown at Figure 8.

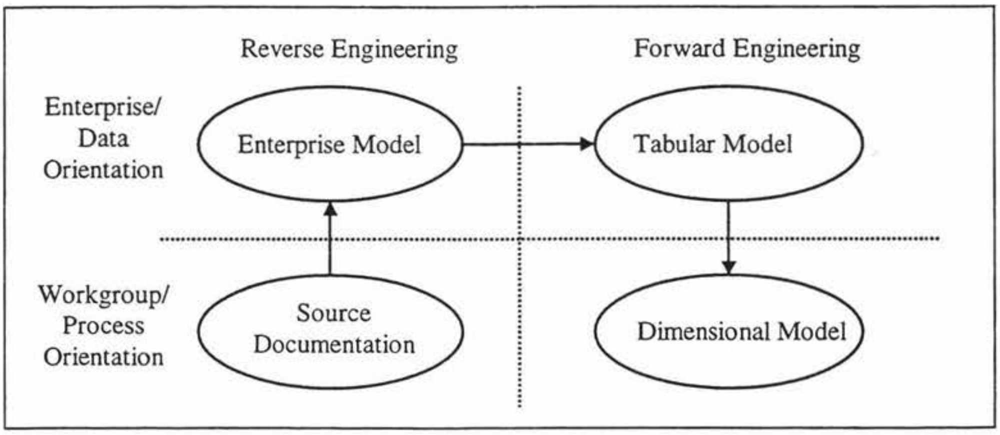


Figure 8: Four models for warehouse design (Winsberg, 1998, p.510).

However, he notes that not all four models may be documented when an isolated data mart is built, and that "it is sufficient to document the source and data mart models" nevertheless "it is important to have a framework even in simple cases, so that future warehouse development is effectively planned" (*ibid.*p.510).

According to Winsberg (1998) the source documentation should reflect the physical data structure of the operational system (this may not be a data model). Whereas the enterprise model "describes all important data in the enterprise at a high level" (p.510). Because enterprise data models projects often fail, Winsberg (1998) recommends the creation of subject area models which are "subsets of the enterprise model that describe a specific part of a business, like customers or products" (p.510). This exercise is useful because it may result in "consistent naming conventions and a list of attribute names and definitions, which are useful in mapping source data to the warehouse" (*ibid.*p.510). The subject area models should be in third normal form and contain no derived or summary data. They can be converted to a tabular model by the following steps: filtering operational data, modelling historic data, modelling changing metadata, introducing derived data, summarising transactions, merging tables for performance, converting column-wise data to row-wise data, restructuring according to data volatility (refer to Winsberg, 1998, p.514).

Subsequent to the creation of the tabular model (or row-wise design), Winsberg (1998) advises the creation of a dimensional model (or star schema). In particular, he suggests following the Kimball (1996) approach which includes: selecting a business process for analysis, identifying numeric measures of performance, or facts about the process, determining important dimensions for the facts, listing the columns that describe each dimension and determining the lowest level of summary stored in the fact table (Winsberg, 1998, p.520).

Indeed Winsberg's (1998) paper is one of the few that describes in detail the data modelling steps, although, like many others, the focus is at the physical level and not at the logical level. This problem is discussed by Date (2000) who provides an excellent critique of the star schema approach. Firstly, he says the approach is ad hoc, and based on "intuition rather than principle" (Date (2000), p.713). The lack of discipline "makes it difficult to change the schema in a proper fashion when (for example) new types of data are added to the database or when dependencies change" (*ibid.* p.713). In addition, star schemas are often created by changing a previous design, which may also have been created by trial and error.

Furthermore, star schemas are really physical models, therefore often no logical design is undertaken. Part of the dilemma is that because the emphasis is on designing for performance, and consequently "decision support systems usually do not distinguish adequately between logical and physical considerations" (*ibid.* p.724). The star schema approach does not always result in a legitimate database design, Date (2000) comments that this problem becomes apparent the more complex the schema becomes.

Another problem is that often designers include different types of facts in one fact table and consequently, "the rows and columns of the fact table typically do not have a uniform interpretation" (*ibid.* p.714). This means that specific columns only apply to specific facts, meaning that the columns not related must allow null values. In addition, the more facts added the more difficult it is to maintain, understand and access.

Problems can also occur within the dimension tables, for example dimension tables can become nonuniform. Date (2000) uses the supplier parts example to illustrate this. For example, when adding total part quantities to the shipments table for each day, quarter and year, other columns may become nonuniform. In this example the columns time period and quantity no longer make sense.

Uncontrolled redundancy may result where designers group quite separate information together in tables that would be better kept separate (in an effort to minimise joins). Sometimes even tables that "simply happen to be accessed together are kept together in the same dimension table" (Date, 2000, p.714), obviously, resulting in "uncontrolled – and probably uncontrollable – redundancy" (*ibid.* p.714). Date (2000) makes it quite clear that he does not agree with Kimball (1996) who says "Efforts to normalize **any** of the tables in a dimensional database solely in order to save disk space [sic] are a waste of time" (p.30).

Inmon *et al.* (1998) suggest the 'corporate information factory' (CIF) as an appropriate architecture for the operational data store (ODS), data warehousing and legacy applications. The role of the data model within this 'common architecture' is made apparent when they comment, "the data model is the data blueprint and intellectually unifies the data warehouse" (*ibid.* p.43). However, this approach is based on the existence and importance of a 'meaningful' corporate data model. In this context the data model should act as a "guide for the ongoing reconstruction of the applications environment" to bring "together the different applications as they are rewritten or as they are modified", and secondly should form the "basis for subject-area design for both the ODS and the data warehouse" (*ibid.* p.43). Like Winsberg (1998) Inmon *et al.* (1998) recommend structuring the ODS or data warehouse towards major subject areas, this orientation "corresponds precisely to the entities that are defined in the high-level logical data model" (p.43). (However, they are actually referring to the physical model – the star schema). While their suggestions are useful for supporting the creation and maintenance of a data model(s) for a data warehouse (or an ODS), they have overlooked the role a data model plays when validating the business rules in the ODS and source systems. In addition, they should have discussed the importance of developing a design data model using methods which result in a model that incorporates the data consumers shared consensual view.

Furthermore, Inmon *et al.* (1998) classify the different types of CIF user as either application users or decision-support system/informational users. They comment that DSS/informational users are "solving or investigating longer-term questions", whereas application users are "concerned with very immediate and very direct decisions" (p.32). From their perspective there are three types of DSS users: 'tourists', 'farmers' and 'explorers'. A 'tourist' specialises in finding a breadth of information, and apparently understands the structure of the CIF and knows where in the structure to find what they require. They remark that, "the tourist is an unpredictable analyst, sort of a (sic) walking directory of information" (*ibid.* p.35). Conversely, the 'farmer' is someone who is predictable and knows what they require before performing a query. The following characteristics are supposed to be typical of the 'farmer': s/he makes use of small amounts of data, is repetitive in the search for information, uses predictable access paths to data and for processing the data, accesses data marts regularly, finds small, interesting patterns, and makes use of presentation tools. The third type of DSS user is known as an 'explorer' who has features similar to the 'farmer' and the 'tourist', but also some unique features, which are: unpredictability and irregularity, analysing large amounts of data, makes farfetched requests, becomes an expert in one area, discovers the relationships between data and uses tools of discovery. Inmon *et al.* (1998) write,

> "Understanding that there are different kinds of DSS users with very different goals and techniques is the first step in resolving many seemingly complex and contradictory facets of the corporate information factory. Without this perspective, many DSS components of the corporate information factory do not make sense" (p.38).

However, not only is it important to understand the DSS audience through discovering their goals and techniques, it is equally important to determine what data they need and in what format. It may be useful to classify the DSS users like this but to also extend this by identifying the different data requirements for each group.

However, in practice this is rarely achieved, as often the end users change roles frequently.

McElreath (1998) provides a description of data modelling approaches for data warehouse design. These are: the star (or dimensional) model, the snowflake model, and the cube model. However, these models are logically equivalent, and are not *different* modelling methods.

Kimball (1996) recommended the dimensional model because it represents a simple data structure. However, since then, Kimball (1998) says the approach has changed somewhat towards the development of a series of data cubes which, when connected form a data warehouse bus architecture. He comments, that the data warehouse market place has begun to mature, "data warehouse owners now have a complete *lifecycle* perspective on their data warehouses" and "the biggest insight that comes from this lifecycle perspective is that each data warehouse is continuously evolving and dynamic" (*ibid.* p.1). Because the environment in which the data warehouse operates is also 'evolving and dynamic', he says expectations and techniques have been adjusted to provide design techniques that are 'flexible and adaptable'. For example, after the business requirements definition stage, this approach includes, the development of a series of data marts to form a data warehouse 'bus' architecture, which will eventually form an overall data warehouse.

Ballard *et al.* (1998) comment, "There is much debate as to which method is best and the conditions under which a particular technique should be selected" (p.36). However there is no definite answer as to which approach is best. They continue, "there are guidelines on which would be the better selection in a particular set of circumstances or is a particular environment"(*ibid.* p.36). Sterling (1998) suggests that using a dimensional approach to data warehouse design is not always appropriate. He comments,

> "If you can implement a dimensional model using views (and indexes and summary tables) on top of a 3NF model, you can get the speed and usability of "canned" queries and many of the tools that insulate users from the underlying tables and drudgery (and errors) of SQL. You can also get the flexibility to extend the model in almost any direction to include purchased or other new data" (p.32).

In contrast Kimball (1996) states,

> "Entity relation data models (sic) are a disaster for querying because they cannot be understood by users and they cannot be navigated usefully by the DBMS software. Entity relation models (sic) cannot be used as the basis for enterprise data warehouses" (p.9).

The pros and cons for each method are discussed by Sterling (1998). Firstly, he claims that normalisation allows the user to ask creative, ad-hoc, queries that extend beyond the preconceived dimensions of a star schema. With a normalised model users can perform ad-hoc queries to discover trends, such as how the weather relates to the sales of particular products. He also comments that by using a dimensional approach, the design is restricted to the dimensions decided upon by the designer.

Sterling (1998) remarks that one of the disadvantages of using a normalised model is the demand on the database, because in the decision support environment users tend to require access to many kinds of facts, so queries of a fully normalised database usually involve many joins. He claims that a large denormalised fact table assists the DBMS to avoid joins, scans, aggregates and sorts, by creating summary data and placing it in the order most likely to be requested. Nevertheless, Date (2000) says that when designers decide to denormalise the database by "prejoining" certain tables, the approach causes as many problems as it solve (pp.696–697). Although, he does agree with the emphasis on designing for performance, but believes that,

> "it should not be allowed to interfere with good design practice. The problem is that, in practice, decision support systems usually do not distinguish adequately between **logical** and **physical** considerations" (*ibid.* p.724).

Sterling (1998) too acknowledges these problems, he writes,

> "Any deviation from the relational logical model, which represents the business mathematically, creates a limitation... Sometimes these trade-offs make good business sense, but often we come to regret them in short order. They make us less efficient and responsive because we have to modify the basic model that we've implemented, instead of just looking at data in another (possibly new) table" (*ibid* p.31).

However, another argument is that the normalised model is difficult for the users to comprehend (Artz, 1997; Mattison, 1997; Sterling, 1998). While this may be true, the users may be shielded from the complexity of the normalised model, by using natural language techniques for verifying the semantic content of the data model. This type of approach is also useful for facilitating user communication (Atkins and Patrick, 2000).

Sterling (1998) also says the complexity of a fully normalised database means during the set up stage a lot more tables have to be created. However, a trained database professional would usually be performing this task, and should not have issue with creating relational tables as the commands to do so are straight forward.

A data warehouse must explicitly consider the temporal aspects of the data it contains, as it usually provides a historical view of an organisation. Time is usually represented as an explicit dimension in every data warehouse, because virtually every data warehouse is a time series. Devlin (1997) says that data warehouse designers have generally had to take a pragmatic approach to temporal issues, such as using timestamps for most of the data. One of the problems with including time in the data warehouse is the issue that often the provision of temporal data in the data warehouse is dependent on the source database. When time is not supported by the source database(s) it may cause difficulties in the data warehouse in terms of managing change over time (such is the case with the organisation studied in this research, they have problems with representing change in the data warehouse especially as records can be physically deleted in the source database). Date (2000) comments that when adding timestamp columns to a key the need for some re-design may occur. In his example, the shipment table is no longer in third normal form once 'month' is added (refer pp.700-701).

The Process of Data Warehouse Design

Although this research is focused on the activity of data modelling for the data warehouse, it is relevant to establish the importance of data modelling as one of the activities in the overall process.

Currently there is no one proven methodology for designing a data warehouse. However, Silverston *et al* (1997) suggest, "The corporate data model is a very good place to start the process of building a data warehouse. It provides a foundation for integration and unification at an intellectual level" (p.239). However, instead *validating* the corporate data model[21] may be a fruitful place to begin the activity of data warehouse design. It is the responsibility of the person creating the data model to ensure that the semantic content of the data warehouse is accurate. This may be achieved through creating a model which may eventually represent information that is meaningful, useful and presented in a way that it can be interpreted easily.

Typanski (1998) claims when selecting a methodology for data warehouse development that the methodology should firstly complement the information architecture and address the needs of the company business personnel. To do this, he says that, the methodology has to emphasise requirements gathering, where the objective is to achieve quality through meeting the requirements. Secondly, the methodology should emphasise data, he comments,

> "Inherent to information engineering is its emphasis on the development of an information system from a data model. This improves the ability of the IS organisation to quickly modify its information systems, adapting them to changing business needs and facilitating the creation of data base designs for the data warehouse" (Typanski, 1998, p.567).

An information engineering methodology combined with object-oriented techniques is recommended, "as the base methodology for an overall information environment that contains a data warehouse" (*ibid.* p.566). (Apart from this recommendation Typanski (1998) does not describe in any detail the steps of this methodology).

Fong and Zeng (1997) describe a decision support life cycle that covers the full life cycle development of a data warehouse. The seven steps of the life cycle are as follows: planning, data requirement (sic) analysing and modelling, analytical database design, data mapping and transformation, data extraction and load, automating data management procedures, application and tool development, and data validation and testing. The second step, "data requirement analysing and modelling" (sic), is described inadequately as, the process where the business needs and data requirements of the users of the system are defined and understood. They suggest

[21] If a corporate data model exists. Otherwise, validating the source data model(s).

possible questions to ask such as: what attributes do the users need? What are the business hierarchies? What data do the users use now and what would they like to have? What levels of detail or summary do the users need? However, often it is difficult to obtain this type of information, as the users may not know this level of detail. Moreover, they provide the following misleading and vague description of data modelling,

> "Data modeling is really the process of translating business concepts into a diagrammatic format that can be converted later into actual database schema. The central focus of data modeling is to provide a logical data model that covers the scope of the development project including relationships, cardinality, attributes and operations" (*ibid.* p.198).

The result of this step is the creation of a 'logical' model which Fong and Zeng (1997) incorrectly term the dimensional model (as noted earlier the dimensional model is not a logical design). They confuse the issue further by describing the third step "analytical database design" (sic), as that which focuses on database design and denormalisation. They comment, "the logical data model resulted from the last step (sic) will be transformed into database schema. ... the star schema design is a simple structure with relatively few tables and well defined join paths" (p.199). Although the star schema may appear to be simple, the join paths are often confusing for users (Date, 2000). Fong and Zeng (1997) describe the database design as that involving the following tasks: designing the schema of the database, denormalising the data, identifying keys, creating indexing strategies and creating database objects. They use the terms star schema and dimensional model in a most confusing way; they infer that the star schema is translated from the dimensional model ('logical' model). However, as Kimball (1996) and Date (2000) point out another name for the dimensional model is the star schema.

Mattison (1996) describes an application-based approach to data warehouse development. However, he does mention that the approach should not be entirely devoid of data modelling. The process of application development is defined as consisting of the following steps: solution development, validation and estimation of the solution, development of a plan of execution and actual system development. Furthermore, to perform these stages effectively he maintains that the following steps are followed:

1. Solution development - A more detailed version of the vision-development process.
2. Data identification - Identification of the data that the solution will require.
3. Data sourcing - Identification of the legacy or external sources of data.
4. Data integrity validation - Validation of the sources of data.
5. Data synchronisation - Determination of how the different sources of data will be synchronised with each other.
6. Back-flush development - Determination of how corrected and cleansed data is going to be fed back into the legacy systems.
7. Data-storage architecture development - Determination of the nature and identity of each of the transitory data stores that will hold the data as it moves from acquisition to the user terminal.
8. Data-transformation mapping - Determination of the specific transformations the data will go through as it moves from the legacy system to the user terminal.
9. Data metrics gathering - Collection of quantitative information about the data.
10. Data modeling - Development of a formal data model for the data.
11. Database design - Development of a physical design for the database.

(Mattison, 1996, pp.127-128)

Mattison (1996) overlooks the need for logical data modelling during data warehouse design and provides misleading advice when he comments,

> "After you have developed the rest of the detail about how the overall data warehouse is going to work, you are ready to begin the modeling step" (p.154).

However, the "detail about the how the overall data warehouse is going to work" may be incomplete if the data modelling activity is deemed unimportant. To explain why the data modelling activity is at step ten, Mattison (1996) makes the comment,

> "...if you want to start off the process of data discipline (sic) with the development of logical data models, then go ahead. Their existence can only help everyone understand the entire process that we have just laid out" (p.155).

Apparently, the intention of this approach is to limit the scope, and to avoid unnecessary involvement in sophisticated and complex data modelling. However, one of the main benefits of data modelling is to define the data required in the data warehouse, thus defining an appropriate scope. Mattison (1996) suggests that the emphasis for the users and developers should be focused on the task at hand: delivering important information to the users as efficiently as possible, however, this can only be successfully achieved by undertaking as Date (2000) says "proper design practice" (p.705).

Interestingly, Date (2000) remarks that database design should always be completed in two stages, logical and then physical. He writes,

> "The rules of logical design do not depend on the intended use of the database – the same rules apply, regardless of the kinds of applications intended. In particular, therefore, it should make no difference whether those applications are operational (OLTP) or decision support applications: Either way the same design procedure should be followed" (p.699).

Summary

Like traditional database design, data warehouse design is a subjective exercise. Moreover, creating a semantically accurate design data model involves a process which is not rigorous, or clear-cut like the normalisation discipline. Despite this, developing a design data model is important for ensuring the physical model implemented is meaningful and useful to the data consumers.

A data warehouse is just a special kind of database (Date, 2000) therefore problems relevant to modelling semantics in traditional databases are also relevant to data warehouse design. While the star schema approach or dimensional model is similar to the relational model, many star schemas are not equivalent to a relationally correct logical design.

Winsberg (1998) describes the data modelling steps necessary for data warehouse design, although, like many others the focus is at the physical level and not at the design or logical level. Date (2000) provides an excellent critique of the ad-hoc star schema approach and in particular describes the problems with such a physical approach.

The findings from this research support the use of a logical design to facilitate the development of a design data model that represents the data consumers consensual view. The ideas discussed throughout this research (specifically in chapter two) are in addition to the more formal semantic features of the relational model.

4 The Research Process

"[A]ll researchers interpret the world through some sort of conceptual lens formed by their beliefs, previous experiences, existing knowledge, assumptions about the world and theories about knowledge and how it is accrued. The researcher's conceptual lens acts as a filter: the importance placed on the huge range of observations made in the field (choosing to record or note some observations and not others, for example) is partly determined by this filter" (Carroll and Swatman, 2000, pp.118-119).

Introduction

This research is at the theory building stage as it is unknown how the semantic data quality of a data warehouse is validated during the development process. According to Shanks *et al.* (1997), "theory building involves developing a set of well defined concepts and their interrelationships" (p.350).

It was important that the research approach selected was appropriate for the exercise of theory building. Shanks *et al.* (1997) utilised a research approach that involved the development of a descriptive process model, although the construction of a model is not the aim of this research, the approach was tailored to incorporate the development of a conceptual framework. The use of conceptual frameworks for theory building is advised by Carroll and Swatman (2000), Miles and Huberman (1994) and Walsham (1993). Figure 9 shows the research approach followed.

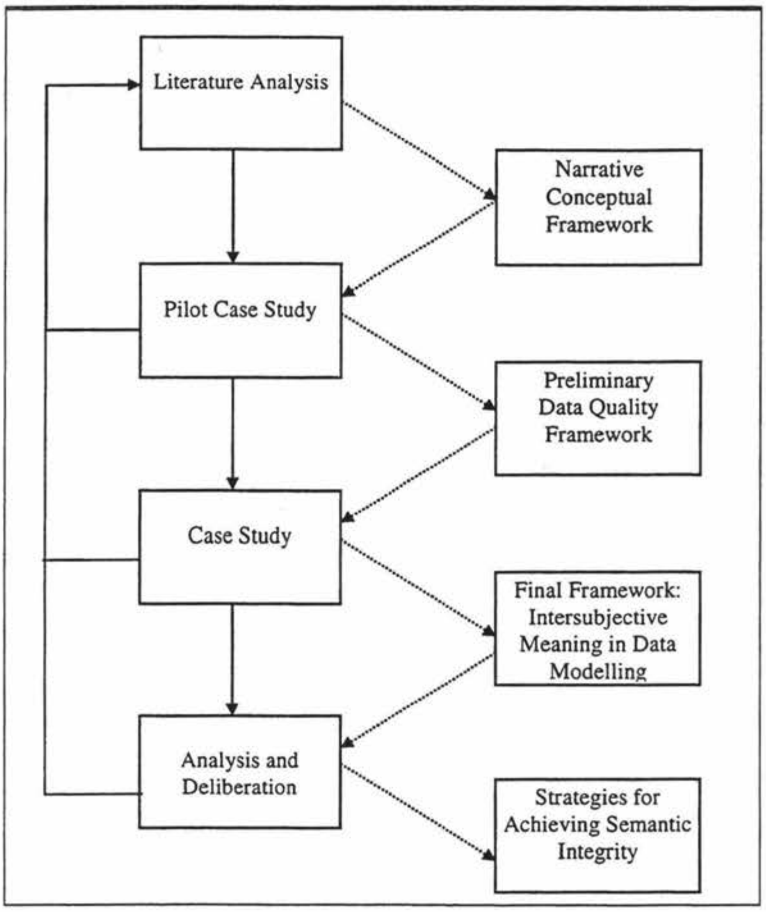

Literature Analysis
Narrative Conceptual Framework
Pilot Case Study
Preliminary Data Quality Framework
Case Study
Final Framework: Intersubjective Meaning in Data Modelling
Analysis and Deliberation
Strategies for Achieving Semantic Integrity

Figure 9: The Research Process (adapted from Shanks *et al.*, 1997)

To devise a framework for achieving semantic integrity in the context of data warehousing, the research approach involved a number of steps. The first step included a literature analysis, and the development of the initial conceptual framework. This framework was narrative and comprised: the philosophical view, the research question, the purpose, and the research propositions (refer to chapter one).

A pilot case study was performed to verify the case study design, and as a result a preliminary semantic data quality framework was devised (see chapter five). Following this preparatory work, the main case study was undertaken. During this process further synthesis and analysis of the literature resulted in the development of the main framework "Intersubjective Meaning in Data Modelling" presented in Chapter two.

The case study data was analysed using N4 (NUD*IST, Non-Numeric Unstructured Data – Indexing Searching Theorising) a qualitative data analysis tool. The decision to use N4 was based on the project's commitment to drawing specific implications based on the original text, and "to develop a comprehensive index to enable themes to be tracked across transcripts" (Cannon, 1998).

Finally, strategies were composed for the generation of meaning from a data model. These strategies address the inhibiting factors for understanding, connotation and intention identified by the participants of the case study (refer to chapter seven).

Rationale

Overall, this research is at the early, formative stage, because there is very little rigorous, systematic research in data warehousing (Shanks *et al.* 1997) and there is a specific lack of research into the activity of data modelling for a data warehouse. Likewise, Atkins (2000) has noted that,

> "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" (p.vii).

Therefore, Benbasat *et al.* (1987) would claim that a case study approach is 'suitable' as the problem is one where "research and theory are at their early, formative stages" (p.369). However, according to Yin (1994), "'how' and 'why' questions are best suited towards case studies, experiments, or histories". Since this research investigates the importance of a validation process for the semantic content of a data warehouse, a detailed case study is appropriate.

In addition, because the philosophical basis for this research is broadly interpretivist, Walsham (1993) would claim that, "the most appropriate method for conducting empirical research in the interpretive tradition is the in-depth case study" (p.14). However, Galliers (1993) discusses four main weaknesses of the case study approach. These weaknesses are: restriction to a single organisation, hence difficulty in generalising, lack of control of variables and different interpretations of events by individuals. The issue with generalisation is accounted for by promoting in this research a type of generalisation that Walsham (1995) describes as "specific implications in particular domains of action" (p.80). Moreover, as discussed in chapters five and seven this research describes generalisations as **tendencies** rather than predictions. The other issues, lack of control of variables and different interpretations by individuals, are accounted for throughout this chapter and the chapters following.

Formulation of the Research Question

The final research question and purpose transpired after several iterations.

Final Iteration:

The main purpose of this research is to explore the importance of semantic integrity during data warehouse design and its impact on the successful use of the implemented warehouse.

Iterations:

1. How data models ~~are~~ utilised during the development of a data warehouse?
2. How are data models used for the purpose of data warehouse development and how are such models validated?
3. The purpose of this research is to investigate the importance of a validation process for the semantic content of a data warehouse and propose a means by which it can be achieved. The question this research will aim to answer is: how are data models validated during data warehouse development.
4. The main purpose of this research is to explore: how the semantic data quality of a data warehouse is validated during the development process. The purpose of the exploration is to consider: the relevance of data modelling to this validation process and ultimately to suggest useful strategies for semantic verification.

Research Propositions:

1. Semantic accuracy is an important critical success factor in determining the effectiveness of a data warehousing project.

2. A 'good' data model is an important critical success factor in determining semantic accuracy.

Data Collection Methods

The primary data collection method for this research was interviewing. This was to ensure best access to the participants interpretations of the data warehouse design (Walsham, 1995, p.78). Other supplementary sources were utilised for example, project documentation such as: the project charter (scope document), the dimension maps and the design documents. Other useful sources of secondary data were: conference papers, publications featuring the case, informal discussions with the participants, application forms (documents from the organisation) and general observations at the site.

Preparation for the data collection was also an important stage, this involved collecting and analysing sufficient background information about each case study site prior to undertaking the interviews. Also all names and positions of all the case participants were obtained. The purpose of this preparation was to provide context for the case study, and to ensure that interview time was only used to obtain information that could not be obtained in any other way (Darke *et al.* 1998). Due to the number of questions asked, and the length of each answer, each interview was tape recorded and later transcribed.

The case study 'database' comprises: the interview tapes (time stamped and participant(s) identified), the raw interview transcriptions (N4 text files), the project unit for the case study in N4, documentation and supplementary information. The contents of the case study 'database' may be accessed through the Department of Information Systems, Massey University, Palmerston North, New Zealand.

Quality Control Measures

To ensure that the data was analysed by more than one researcher the transcriptions and coding analysis in N4 were made available to the research supervisor, this was to confirm the assignation of appropriate meaning (Galliers, 1993; Greenhalgh, 1997).

The case study follows the principle of data triangulation where some of the findings are supported by another researcher (Mair, 1999), who also performed a case study at the same organisation. The purpose of data triangulation is to reduce bias by providing multiple instances from different sources. Moreover, as mentioned in the previous section the 'evidence' for the case studies was derived from multiple stakeholders.

However, the entire research process was carefully guided by the creation of guidelines for single study case research. The purpose was to ensure quality in this research process, and also to assist other researchers undertaking similar case study research. Creating the guidelines was fundamental for ensuring quality in this research because it is evident from the literature that there is no unified list of criteria for case study research. Therefore, from the literature and the experience gained from this research, a set of practical guidelines to assist an active researcher are described.

Summary

The rationale for choosing a single case study approach was mostly to provide the setting for a detailed investigation, and a suitable environment for rich description and understanding (Walsham, 1995). As a result of both the pilot case study and the literature analysis, the research question was refined. The research process followed throughout this study was influenced by the work of Shanks *et al.* (1997) and Carroll and Swatman (2000). The development of a conceptual framework (Miles and Huberman, 1994, Carroll and Swatman, 2000) provides the theory building mechanism for this research (refer chapter 2).

The main method for collecting the data was through interviewing, although considerable secondary data was collected, such as: conference papers, publications featuring the case, informal discussions with the participants, documents from the organisation and general observations at the site. The data was analysed by using the qualitative data analysis tool N4. To assist with the quality control of the research method a set of guidelines for single case study research have been devised, these are described in the next chapter.

5 Practical Guidelines for Single Case Study Research

"Interpretive research can help IS researchers to understand human thought and action in social and organizational contexts; it has the potential to produce deep insights into information systems phenomena including the management of information systems and information systems development. As the interest in interpretive research has increased, however, researchers, reviewers, and editors have raised questions about how interpretive field research should be conducted and how its quality can be assessed" (Klein and Myers, 1999, p.67).

Introduction

In the previous chapter, it was concluded that a single case study approach is most suitable for this research (Benbasat *et al.,* 1987; Walsham, 1993; Yin, 1994). To guide the research process described in chapter four, guidelines have been created for single study case research. The purpose was to ensure quality in this research process, and to also assist other researchers undertaking similar case study research. Creating the guidelines was important for ensuring quality in this research because there is no such list of criteria for case study research. Therefore, from the literature and experience gained from this research, a set of practical guidelines to assist an active researcher have been devised.

Previous Research

Case study research may be classified as a qualitative method, where the purpose of which is to try to understand, or interpret, phenomena in terms of the subjective meanings people bring to them (Denzin, 1994).

The work of Yin (1994) on case study research design is recognised and cited by many IS researchers as providing an important contribution to case study design. According to Yin (1994), a case study is defined as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" (*ibid.* p.13). However, the nearest the author comes to addressing the challenge of 'quality' in case study design is through the use of a case study protocol. He quite correctly deduces that: "A case study protocol is more than an instrument. The protocol contains the instrument but also contains the procedures and the general rules that should be followed in using the instrument." (*ibid.* p.63). He maintains that the protocol should have the following sections: an overview of the case study project, field procedures, case study questions, and a guide for the case study report. All of these are important but the problem is that it often difficult to accurately document this level of detail at the beginning of the research project. The protocol steps described by Yin (1994) would be of particular use for IS case study research if they were components of a more detailed set of practical guidelines. This research will provide that practical set of guidelines for conducting case study research.

Darke *et al.* (1998) provide an excellent discussion on the pragmatics of conducting case study research. They address five difficulties when undertaking case study research: selecting appropriate research, designing shaping and scoping a case study research project, obtaining the participation of organisations, collecting case study data from case participants and establishing rigour in writing up case study research. Also, situations where a case study research method may **not** be appropriate are described, these are:

> "where a phenomena is well-understood and mature, where constructs exist already and are well-developed, where understanding of how and why the particular phenomenon occurs is not of interest, and where understanding of the contexts of action and the experiences of individuals in single settings is not relevant" (*ibid,*p.280*).*

Similarly, Klein and Myers (1999), have devised a set of principles for conducting and evaluating interpretive field studies in IS. The principles they propose are fundamental ideas derived from philosophical writings that may be helpful to authors and reviewers. Although their principles apply mostly to the conduct and evaluation of interpretive research in the hermeneutic nature, they nevertheless provide valuable input for justifying the guidelines proposed in this research. Most importantly, they conclude that,

> "while not all of the principles may apply in every situation, their systematic consideration is likely to improve the quality of future interpretive field research in information systems (especially that of a hermeneutic nature)" (*ibid.* p.70).

Miles and Huberman (1994) describe their book "Qualitative Data Analysis" as a source book bringing together a collection of resources. They have synthesised the literature on qualitative data analysis prior to 1994, and also incorporated the work of other active qualitative researchers in their work. Their work provides a very detailed description of what qualitative data analysis involves. They devote six chapters to define the three concurrent activities of analysis: data reduction, data display and conclusion drawing/verification. The scope of their work covers both single study and multiple study cases.

Like Yin (1994) the research of Walsham (1995) is recognised and cited by many as providing an important contribution to the nature and method for interpretive case studies. He also recognised the lack of "a synthesised view of the nature and conduct of case studies with specific reference to the field of computer-based IS" (*ibid.* p.74). Walsham (1995) discusses: the research tradition of interpretive research, the use of theory in interpretive studies, four types of generalisations (extending Yin's work), and conduct of empirical work.

Most recently, Carroll and Swatman (2000) have provided an integrated methodological framework for building theory in information systems research called Structured-case. They present a research cycle that describes: the research plan, data collection, data analysis and reflection, and using a conceptual framework they speak of "a spiral towards understanding" (p.120).

Determining Practical guidelines for Case Study Research

Greenhalgh (1997) proposes nine questions for evaluating papers that describe qualitative research (see table 1). She has devised questions (or criteria) based on her own research and teaching experiences, and from the research of Denzin and Lincoln (1994), Mays and Pope (1996) and Britten *et al.* (1995).

1	Did the paper describe an important clinical problem addressed via a clearly formulated question?
2	Was a qualitative approach appropriate?
3	How were the setting and the subjects selected?
4	What was the researcher's perspective, and is this been taken into account?
5	What methods did the researcher use for collecting data, and are these described in enough detail?
6	What methods did the researcher use to analyse the data , and what quality control measures were implemented?
7	Are the results credible, and if so, are they clinically important?
8	What conclusions were drawn, and are they justified by the results?
9	Are the findings of the study transferable to other clinical settings?

Table 4: Nine guidelines for evaluating papers (Greenhalgh, 1997)

These questions provided the groundwork for the guidelines which have emerged from this research. Indeed, these questions are very appropriate for developing case research guidelines, although their original intention was for evaluating papers that describe qualitative research in health care.

Subsequently, through an analysis and synthesis of prominent literature on case study design and assessing qualitative research, practical guidelines to assist an active researcher have been extracted. While recognising that cases studies are not necessarily qualitative, (Stake, 1994) the guidelines we present are focused on qualitative case studies.

The guidelines as presented in Table 5 do not reflect the order in which they usually transpire. The objective was not to provide a regimented list of rules, but to provide suggestions or advice for an active, case study researcher.

To clarify the guidelines and formalise the research process, a framework was utilised for classifying the process of case study research. The framework proposed by Bronts *et al.* (1995) of IS development methods describes components which are useful and appropriate for classifying the guidelines. This framework captures the essence of the guidelines suggested. The components of the framework are: way of thinking, way of working, way of controlling, way of supporting and way of communicating. However, the component 'way of modelling' is not appropriate and has been excluded. The *way of thinking* describes the assumptions and viewpoints of the researcher in the context of current research, and thus the philosophical framework of the research is made explicit. The *way of working* - defines and orders the tasks and sub-tasks that are to be performed in the research exercise, and also provides guidelines and heuristics on how these tasks should be carried out. The *way of controlling* - sets out how the research exercise should be managed. The *way of supporting* details how tools, support the research exercise. The *way of communicating* describes the form in which the research is to be communicated.

These guidelines cover both the research approach (or strategy) which is "a way of going about one's research, embodying a particular style and employing different methods" and the research method which is "a way to systemise observation, describing ways of collecting evidence and indicating the type of tools and techniques to be used during data collection" (Cavaye, 1996, p.227).

Practical Guidelines for Case Study Research

The following table is the list of guidelines classified according to the framework of Bronts *et al.* (1995). The guidelines have emerged through a synthesis of the work by: Carroll and Swatman (2000), Darke *et al.* (1998), Greenhalgh (1997), Klein and Myers (1999), Maxwell (1996), Miles and Huberman (1994), Patton (1990), Richards (1997), Walsham (1995), Yin (1994) and Zinatelli and Cavaye (1992).

Component	Guideline	Authors
Way of thinking	Provide an argument for why a case study is appropriate.	Greenhalgh, Darke *et al,* Zinatelli and Cavaye.
	State your philosophical stance and perspective. Take any bias into account when performing data analysis.	Walsham; Klein and Myers.
Way of controlling	Define and use some form of quality control measures.	Greenhalgh, Miles and Huberman, Yin.
	Ensure that the results are credible, and important to IS practitioners.	Greenhalgh, Moody and Buist.
	Determine how to draw conclusions and justify the results through the appropriate use of theory.	Walsham.
Way of working	Construct a clearly formulated question that describes an important IS issue or problem of interest.	Greenhalgh, Yin, Darke *et al.*
	Create a first cut conceptual framework	Miles and Huberman, Carroll and Swatman.
	Devise first cut case study questions.	
	Make explicit the research approach.	Shanks *et al.*
	Perform a pilot case study	Yin.
	Determine criteria for selecting the appropriate case and participants.	Greenhalgh, Patton, Maxwell.
	Refine the case study questions based on lessons learnt from the pilot study.	
	Revisit your research purpose/question and modify the conceptual framework as necessary.	Greenhalgh, Klein and Myers, Miles and Huberman, Carroll and Swatman.
Way of supporting	Choose appropriate methods for collecting data. Ensure that these are described in enough detail.	Greenhalgh, Walsham.
	Employ a systematic way to analyse the data.	Greenhalgh, Richards, Miles and Huberman.
Way of communicating	Create a plan for the final report.	Yin, Walsham.
	Determine how the case study findings might be transferable to other settings.	Greenhalgh, Miles and Huberman.
	Determine how to present the case study findings to the academic and practitioner communities.	Darke *et al.*, Miles and Huberman.

Table 5: Guidelines for Case Study Research

Way of thinking

Provide an argument for why a case study is appropriate.

This guideline requires the researcher to not only provide an explanation of case study research, but to also provide justification for choosing the approach. This should involve defining the strengths and weaknesses of case studies (see Yin 1994, Zinatelli and Cavaye 1992). The methodology story should also tell the audience if the research was successful or not.

The predominant argument for undertaking case study research is one where the phenomena of study cannot be separated from its natural setting, and the researcher wishes to cover contextual conditions. In the heath arena, Greenhalgh (1997) suggests that if the objective of the research is to explore, interpret, or obtain a deeper understanding of a particular clinical issue, then a qualitative method is the most appropriate. Yin (1994), adopts an implicitly positivist approach to case study research. He argues that: 'how' and 'why' questions are best suited towards case studies, experiments, or histories". Furthermore, Benbasat *et al.* (1987), claim that case studies are suitable for problems where "research and theory are at their early, formative stages" (p369).

State your philosophical stance and perspective, take any bias into account when performing data analysis.

It is important that the researcher reflects on her/his philosophical stance and states it clearly when writing up their work. The main reason in stating this up front is because it affects every aspect of the research process, from how the evidence is collected to how the results are interpreted. There are a multitude of papers and research discussing the positivist and interpretivist philosophical traditions (Atkins 2000; Darke *et al.* 1998; Klein and Myers 1999; Travis 1999; Walsham 1995).

It is improbable that case study research could be conducted by a researcher with no views at all and no ideological or cultural stance. Because of this problem, it is imperative that the researcher describes in detail where they are coming from so that the results can be interpreted accordingly (Greenhalgh, 1997).

The principle of suspicion as described by Klein and Myers (1999) "requires sensitivity to possible 'biases' and systematic "distortions" in the narratives collected from the participants". They illustrate this with an example by Forrester (1992) whose approach went "beyond understanding the meaning of the data" to that of reading "the social world behind the words of the actors".

Way of Controlling

Define and use some form of quality control measures.

Quality control methods as described by Greenhalgh (1997) are: ensuring that the data has been analysed by more than one researcher "to confirm that they are both assigning the same meaning to them". They do however mention that this is often difficult to achieve, "we could find no data on the interobserver reliability of any qualitative study to illustrate this point".

Miles and Huberman (1994) comment that through the triangulation of data a finding can be supported by showing that independent measures of it agree with it, that is bias can be reduced by providing multiple instances from different sources. Likewise, Yin (1994) says evidence for case studies should come from at least six sources. These sources may be: documentation, archival records, interviews, direct observations, participant-observation and physical artifacts (*ibid.* pp. 78-99). Greenhalgh (1997) also agrees when she suggests that a good qualitative study will use more than one research method.

Other quality control methods such as the creation and maintenance of a case study 'database' is recommended by Yin (1994). He says the creation of a case study database is very important (see pp.94-98) and claims that the lack of a formal database for most case study efforts is a major shortcoming of case study research. Furthermore, it is important to demonstrate a chain of evidence to increase the reliability of information in a case study evidence (Yin, 1994). This may be achieved by cross-referencing supporting documents during the data collection and data analysis phases, and creating an annotated bibliography of documents (Darke *et al.* 1998).

Determine how to draw conclusions and justify the results through the appropriate use of theory.

Eisenhardt (1989) discusses three distinct roles theory can play in organisational research: theory which guides data collection and analysis, theory that emerges as an iterative process of data collection and analysis and theory as a final output of the research. Walsham (1995) illustrates the roles of theory with examples and warns of the danger of using theory to guide data collection and analysis,

> "there is a danger of the researcher only seeing what the theory suggests, and thus using the theory in a rigid way which stifles potential new issues and avenues of exploration" (p.76).

As mentioned earlier Carroll and Swatman (2000) have proposed a methodological framework for building theory in information systems called Structured-case. Their framework describes a process model made up of three structured components: a conceptual framework, a pre-defined research cycle and a literature-based scrutiny of the research findings. Underpinning the Structured Case methodology is the ability of the researcher to create a formal conceptual structure at the outset of the research. They agree with the definition of 'conceptual framework' as provided by Miles and Huberman (1994) however extend this when they comment, "The conceptual framework is the researcher's *representation* of the conceptual structure brought to the research process" (p.118).

Like Miles and Huberman (1994), Carroll and Swatman (2000) suggest that the conceptual framework is formed from "the research themes, existing knowledge about which is gathered from the literature and insights, filtered by a researcher's theoretical foundations" (p.118). Therefore the conceptual framework and subsequent revisions of it "documents both the *process* through which the theory was built and its *links* to the data collected in the field" (*ibid.* p.121). They emphasise that their approach is particularly suited to building theory of the middle range that involves some level abstraction but is closely linked to observations.

Carroll and Swatman (2000) describe theory building in IS research and recommend the creation of a conceptual framework. However, Walsham (1993) discusses the limitations of using frameworks when he comments,

> "a researcher should have an analytic framework, but should retain a degree of scepticism concerning its value... a theory is a way of seeing and a way of not-seeing, since the use of a particular theory excludes other ways of viewing the same events" (p. 70).

Four types of generalisations are discussed by Walsham (1995): the development of concepts, the generation of theory, the drawing of specific implications and the contribution of rich insight. These should be used as a basis for theorising case study research. He asserts that the generalisations should be seen as,

> "explanations of particular phenomena derived from empirical interpretive research in specific IS settings, which may be valuable in the future in other organisations and contexts" (*ibid.* p.79).

He also stresses that the generalisations are not necessarily mutually exclusive and are further described as tendencies rather than predictions. Each generalisation type is illustrated by an example in his work. Elsewhere, Walsham (1993) suggests that the validity of a generalisation from an individual case or cases depends on the "plausibility and cogency of the logical reasoning used in describing the results from the cases, and in drawing conclusions from them" (p.15).

An active researcher should determine the type of generalisation relevant to their research goal and research strategy (using Walsham's (1995) descriptions). For example, case studies may be used to develop concepts, or to generate a theory by integrating several concepts, propositions and world-views. However, the type of generalisation derived is dictated by the number of cases to be studied. Darke *et al.* suggest, "single cases provide for in-depth investigation and rich description. Multiple case-designs allow literal or theoretical replication and cross-case comparison". (The scope for this research has been restricted to providing guidelines for single-study cases).

The research of Walsham and Waema (1994) presents an example of how specific implications are drawn, based on an in-depth case study of IS development in a financial services company. The generalisation of 'rich insight', is designed to, "capture insights from the reading of reports and results from case studies that are not easily categorised as concepts, theories or specific implications"(*ibid.* p.80). The purpose of his discussion is to classify research that generates broader and more diffuse implications.

A useful way to justify the research conclusions is by asking the following questions: how comprehensible would this explanation be to a thoughtful participant in the setting? and how well does this explanation cohere with what we already know? (Mays and Pope, 1996).

Ensure that the results are credible and important to IS practitioners.

The aim of this guideline is to ensure results obtained from case study research are both credible and practical. Greenhalgh (1997) discusses the issue of assessing credibility in qualitative research, "It often takes little more than plain common sense to determine whether the results are sensible and believable and whether they matter in practice" (p.160). She also emphasises that the researcher cites actual data. The results should be "independently and objectively verifiable" to do this the researcher must ensure that all quotes and examples are indexed so they can be "traced back to an identifiable subject and setting" (*ibid.* p.160). The beauty of using a qualitative data analysis tool is that it provides a formal index system, which identifies the source of all quotes.

Way of Working

Construct a clearly formulated question that describes an important IS issue or problem of interest to be researched.

Researchers should try to determine the perceived value of the research. This may include determining the potential usage. For example, Moody and Buist (1999) highlight this when they comment,

> "the value of information is dependent not on its volume but its *usage* - the more it is used, the more value can be extracted from it. Information which is never used has zero value" (p.651).

They go on to comment that research knowledge in IS is only valuable if it contributes to more effective use of information technology in practice. However, they have not emphasised the importance of research from a pedagogical standpoint.

The aim of this guideline is to remind the researcher that not only do we have to formulate a precise research question, but to also research an important IS issue. Darke *et al.* (1998) support this when they comment,

> "it is important to ensure that the questions are appropriate in terms of their interest, significance and value for both the research and practitioner information system communities" (p.280).

They also suggest that the research question should be one that can be answered in a useful way. Thereby, the research question should state what is to be discovered, whereas hypotheses should provide the initial answer(s) to the question. Miles and Huberman (1994) suggest that many researchers explicitly state their ideas as part of the process of theorising and data analysis, they refer to this as generating propositions rather than hypotheses. In qualitative research hypotheses are usually developed after the researcher has begun the study, as Maxwell (1996) comments, "hypotheses in qualitative research... are grounded in the data and are developed and tested in interaction with it, rather than being prior ideas that are simply tested against data" (p.53).

Create a first cut conceptual framework

A conceptual framework explains, "either graphically or in narrative form, the main things to be studied - the key factors, constructs or variables - and the presumed relationships among them" (Miles and Huberman, 1994, p.18). One of the main motivations for developing a preliminary conceptual framework is to help focus the research and avoid 'information overload'. Carroll and Swatman (2000) agree with Miles and Huberman (1994) and comment that, "all researchers bring some kind of conceptual structure to the research process" and that they,

> "interpret the world through some sort of conceptual lens formed by their beliefs, previous experiences, existing knowledge, assumptions about the world and theories about knowledge and how it is accrued" (p.118).

According to their approach a formally defined conceptual structure is imperative throughout the entire research process. The initial conceptual framework is revised many times until the point of closure, and in some cases may change significantly (as was the case in this research).

Devise first cut case study questions.

After formulating the research question and performing a literature review, it is appropriate to start thinking of possible interview questions. It is important to start devising interview questions early on, as they help to focus the research. The questions may be fairly broad, and may remain so until the pilot case study is completed.

Although a discussion on the research literature is outside the scope of this paper, Darke *et al.* (1998) comment that, the literature review should, "provide a basis for careful design of the research project structure and scope so that an appropriate unit of analysis and number of cases can be determined".

Once the participants have been determined, a useful guide is to group questions according to their role. For example, organising questions into categories such as: participant background, general, methodology, data modelling techniques and project outcome, may be helpful when devising questions to ask the data manager, the data modeller and the data consumers. However, care must be taken to ensure that only relevant questions are asked of the participant.

Another technique is to devise questions based on the conceptual framework developed earlier. This process may also involve changes to the conceptual framework (see Miles and Huberman, 1994. pp. 22-25). However, the development of research questions may precede or follow the construction of a conceptual framework (Miles and Huberman, 1994).

Make explicit the research approach

As defined earlier the research approach (or strategy) is the particular style and methods used for going about the research (Cavaye, 1996). The purpose of this guideline is to make sure the approach and techniques for data collection and analysis are described in detail (including the rationale for their selection). For example, Shanks *et al.* (1997) described their approach pictorially (see p.351) and explicitly defined each component.

Carroll and Swatman (2000) recommend,

> "Researchers should carefully match the research approach to the research topic, situation and available resources, rather than sticking to just one approach" (p.116).

Through the research cycle of structured-case, this challenge should be addressed.

Perform a pilot case study

A pilot case study can be viewed as the 'dress-rehearsal' of the final case study. Although not often admitted, performing a pilot case study is a very useful method to ensure the interview questions are appropriate and useful for the purpose of extracting the required the information. Refer to chapter five of this report for a detailed description of the pilot case study purpose.

Determine criteria for selecting an appropriate case study and stakeholders

This guideline emerged from the work of Patton (1990), Greenhalgh (1997) and from this particular research experience of determining appropriate criteria for selecting data warehousing projects.

The researcher should conduct an intentional selection process to choose specific settings, persons or events (Patton, 1990). Likewise, Greenhalgh (1997) remarks that to gain an in-depth understanding of the participants experience we should, "deliberately seek out individuals or groups who fit the bill" (p.157).

In our experience, the process for project selection involved a number of steps:

1. defining the unit of analysis
2. determining the stakeholders

3. precisely defining the project criteria

4. verification of the project criteria by an external party and the research supervisor

Step one can be found in any book discussing case study research design (e.g. Yin, (1994). The point to note here is that not only should the researcher's case study questions coincide with the research purpose, but the unit of analysis must also. For example, if aim of the research is to look at, "the importance of data models in database design", then, the unit of analysis, the organisation, may include: the project manager, the data modeller(s) and the end user(s). This may help focus the data collection towards to the research question and propositions. Likewise, Giannoccaro *et al.* (1999) undertook a case study where they interviewed four stakeholder types: data producers, data custodians, data consumers and data managers. Darke *et al.* (1998) also agree when they comment, "The unit of analysis must also provide for sufficient breadth and depth of data to be collected to allow the research question to be adequately answered" (p.280).

Step two requires an identification of criteria to seek out appropriate individuals or groups. Once again, this step is described in many source books (Yin 1994; Miles and Huberman 1994). However, these texts fail to mention how to verify the criteria once they have been defined. Ideally, the researcher should aim to consult at least two sources for verifying the criteria. For example, a practitioner (in the relevant area) and the research supervisor. Obviously, this is to prevent the researcher from choosing inappropriate criteria. Criteria verification may provide much needed feedback (especially practical) early on, as to the potential usefulness of the research.

Refine the case study questions based on lessons learnt from the pilot study.

Often, the list of interview questions is modified and additional questions added, after the pilot case study. Indeed, other changes may be required, for example, a different type of participant may need to be interviewed. This amendments process is the formalisation of the case study questions and logically leads onto the next guideline, 'revisit the research question'. Through this process, it becomes evident how important the 'dress-rehearsal' or pilot case study was.

The pilot case study may also uncover any difficulties the participant(s) have in terms of understanding the concepts/questions asked of them. Therefore, to avoid misunderstandings it may be useful to include a glossary of terms in the initial letter sent to participants. This may help to ensure that the participants understand what they are being asked. Pilot case studies may also necessitate a change in the order of certain key questions, and the timing of discussion about the setting. For example, a discussion about the organisation (or setting) to provide context for the interview is useful at the beginning of the interview as it can lead to further discussion.

Revisit the research purpose/question and modify the conceptual framework as necessary.

Changes are acceptable and welcomed if this means the researcher can clarify her/his question further. Undertaking a pilot case study is a really useful technique for refining the research question. Greenhalgh (1997) legitimises modifying the research question (or hypothesis) as these types of changes may show sensitivity to the richness and variability of the subject matter.

Elsewhere, Klein and Myers (1999) discuss the importance of dialogical reasoning. This principle "requires sensitivity to possible contradictions between the theoretical preconceptions guiding the research and design and actual findings with subsequent cycles for revision". They also stress that the researcher should make the historical intellectual basis of the research as clear as possible. This principle is illustrated with three examples (Orlikowski 1991; Walsham and Waema 1994; Myers 1994), although they do suggest that the examples are weak in terms of explicitly discussing the dialogical aspect of the research. Therefore, to help ensure the intellectual basis and the dialogical aspect of the research is made clear the researcher should not only refine his/her research question, but explain and record the reasons for these changes.

Way of Supporting

Choose appropriate methods for collecting data. Ensure that these are described in enough detail.

Walsham (1995) says that interviews should be the primary data source for interpretive case studies,

> "since it is through this method that the researcher can best access the interpretations that the participants have regarding the actions and events which have or are taking place" (p.78).

However, Yin (1994) says that at least six sources for collecting data should be accessed, such as: documentation, archival records, interviews, direct observations, participant-observation and physical artifacts. Useful factual information can be obtained through examining annual reports or by obtaining written answers to structured questions (Darke *et al.* 1998). Internal magazines and organisational bulletins may be used to supplement information gained through other sources (*ibid.* p.282).

Preparing for data collection is also vitally important when undertaking case study research. Sufficient background information about a case study site should be collected and analysed. Also all names and positions of all potential case participants should be obtained and interview time should only be used to obtain information that cannot be obtained in any other way (Darke *et al.* 1998). (For techniques on how to collect case study data refer to Darke *et al.* 1998, Walsham, 1995 and Yin, 1994).

A researcher must provide a detailed account of the research procedure including the data analysis methods. One motivation for ensuring sufficient detail of the research procedure is provided, is in the event that another researcher wishes to replicate the study. Indeed, describing the chosen data collection methods is vitally important for qualitative research because, "It may have to be lengthy and discursive since it is telling a unique story without which the results cannot be interpreted" (Greenhalgh 1997, p.159). To ensure adequate detail is recorded about the data collection methods the researcher should ask the following questions: have I provided enough information about the methods used? Are these methods a sensible and adequate way of addressing the research? (*ibid*, p.159).

Employ a systematic way to analyse the data.

Richards (1997) provides a commentary on the reasons to support the use of computers for analysing qualitative data. She comments that the main reason for using a qualitative data analysis tool is to enable access to large quantities of unstructured qualitative data. The latest qualitative data analysis tools, such as NUD*IST or ATLAS.ti will get the researcher to the data and then provide context. However, a prerequisite for using such a tool is the need for a 'thinking' researcher who has a sense for what they are trying to do. Indeed, such tools will not analyse the data, or 'see' patterns in the data, but they are useful for managing and presenting qualitative data. Therefore, despite the existence of computer software for analysing qualitative data, researchers still require a knowledge data analysis techniques, in particular a researcher requires an understanding of coding methods.

Miles and Huberman (1994) describe the activity of coding as analysis, they mention three types of codes: descriptive, interpretive and pattern codes. A descriptive code "attributes a class of phenomena to a segment of text", whereas codes that are devised as the researcher becomes more knowledgeable about the case are interpretive, for example a descriptive code labelled 'motivation' may be split into two interpretive codes of 'private motivation' and 'public motivation'. A coded segment of interview notes "illustrates an emergent leitmotiv or pattern that you have discerned in local events and relationships" (*ibid*, p.57).

Pattern codes are those that emerge as a result of subsequent interviews, pattern codes are added to earlier codes to indicate the inferred theme or pattern. Therefore, according to this classification, codes reflect the different levels of analysis from descriptive to the inferential. Codes can also occur at different times in the analysis, usually starting with the more descriptive codes and the more inferential codes follow. However, they say the most important point to note is that "codes are astringent - they pull together a lot of material, thus permitting analysis". Pattern codes should signal "a theme that accounts for a lot of other data" and should suggest thematic links, by "grouping disparate pieces into a more inclusive and meaningful whole" (*ibid*, p.58).

They also discuss three different approaches for coding: a priori, inductive and accounting-scheme guided. The first approach involves creating a start list of codes prior to fieldwork. Miles and Huberman (1994) promote this approach because the list should emerge from the conceptual framework, list of research questions and propositions of the study. They also discuss the appropriateness of the 'grounded' approach of Glaser and Strauss (1967) especially for exploratory research. When using this approach they comment "The analyst is more open-minded and more context-sensitive, although, here too, the ultimate objective is to match the observations to a theory or set of constructs" (Miles and Huberman, 1994, p.58).

The third approach is where a general accounting scheme for codes is created that is not content specific, but "points to the general domains in which codes can be developed inductively" (*ibid*.p.61). Apparently, these schemes help the researcher to think about categories in which codes will be developed. They provide two examples of accounting-scheme approaches to coding, Lofland (1971) and Bogdan and Biklen (1992). The latter divide codes using the following categories:

setting/context, definition of the situation, perspectives, ways of thinking about people and objects, process, activities, events, strategies, relationships/social structure and methods.

Miles and Huberman (1994) comment that not all categories have to be included in any particular study, only those deemed appropriate. The most important point to glean from this is the importance of structure, where the choice of approach is less important than whether a conceptual and structural order is used. They comment that codes should "relate to one another in coherent, study-important ways; they should be part of a governing structure" (*ibid.*p.62). From this it can be deduced that, the coding structure must relate and be based upon the developing conceptual framework and research question.

Other techniques described by Darke *et al.* (1998), are: content analysis, conversation analysis and discourse analysis. (These are also discussed by Miles and Huberman, 1994).

Way of Communicating

Create a plan for the final report

Walsham (1995) asks the question - "what should be reported in an interpretive case study?" He replies for the collection of field data: details of the research sites chosen, the reasons for this choice, the number of people who were interviewed, what hierarchical or professional positions they occupied, what other data sources were used, and over what period was the research conducted. For data analysis he replies, how the field interviews and other data were recorded, how they were analysed and how the iterative process between field data and theory took place and evolved over time.

Throughout the conduct of the case study the researcher must dedicate some time to focusing on the design of the final report. As Yin (1994, p. 73) points out this is because there is no uniformly acceptable outline for the formatting of case study reports, unlike other research strategies. Determining the plan for the final report could take the form of an initial structure or outline for the research. This outline should also contain an annotated bibliography section for any supporting documentation. Often the case study plan or research outline will change as a result of the data collection.

As mentioned earlier to increase the reliability of information in a case study it is important to demonstrate a chain of evidence. The structure of the report should support this requirement by linking each chapter of the report eloquently. This would not be the case if the structure was merely composed of: the research problem, conceptual framework, methodology, data analysis, conclusions and discussion. A circular linkage needs to be provided between the research questions, methodology, data collection and interim analyses (Miles and Huberman, 1994, p.298). In each of these chapters the researcher should provide history and context by linking back to the research purpose.

Determine how the case study findings might be transferable to other settings.

Greenhalgh (1997) points out that one of the most common criticisms of qualitative research is that often the findings are only applicable to the limited setting in which they were obtained. Essentially, the conclusions of a study should be transferable to other contexts. She suggests that the "use of a true *theoretical* sampling frame greatly increases the transferability of the results over a "convenience" sample" (*ibid.* p.161).

Miles and Huberman (1994) provide a list of twelve queries for a researcher to usefully ask when considering external validity/transferability/fittingness. For the purposes of single study cases we have extracted the following queries from their original list:

1. Are the characteristics of the original sample of persons, settings, processes fully described enough to permit adequate comparisons with other samples?
2. Do the findings include enough "thick" description" for readers to assess the potential transferability, appropriateness for their own settings?
3. Are the processes and outcomes described in conclusions generic enough to be applicable in other settings, even in ones of a different nature?
4. Is the transferable theory from the study made explicit?
5. Does the report suggest settings where the findings could be tested further?

(Miles and Huberman, 1994, p.279).

Determine how to present the case findings to the academic and practitioner communities.

The case study should be reported in a *useful* and accessible form to academics and practitioners if we are to help control the volume of information currently being produced. Not only should research knowledge contribute to more effective use of information technology in practice, but should also provide a valuable pedagogical resource. This may require the generation of more than one type of paper from the research depending on the intended audience.

Critical Appraisal Guidelines

The guidelines of Table 5 were used in the development of critical appraisal guidelines for single case study research. The purpose of which is to provide a means to validate both academic and practitioner sourced literature, through the use of hierarchies of evidence. Refer to appendix five for the critical appraisal guidelines devised from these guidelines, the research of McKay and Marshall (2000) and Wheeler (2000).

Summary

This chapter presents guidelines for undertaking single case study research. This was an important exercise because no existing list of criteria for single case study research exists. The exercise involved synthesising the existing literature on case study design and applying the guidelines of Greenhalgh (1997) for assessing qualitative health care research.

The guidelines were classified according to the framework by Bronts *et al.* (1995), which grouped the guidelines according to: the 'way of thinking', the 'way of working', the 'way of controlling', the 'way of supporting' and the 'way of communicating'.

Overall, the guidelines are to help ensure quality in this research process and to also assist other researchers undertaking similar case study research. Furthermore, the guidelines were one input in the process of creating critical appraisal guidelines for single case study research.

6 The Dress Rehearsal

"The so-called iterative approach (altering the research methods and the hypothesis as you go along) used by qualitative researchers shows a commendable sensitivity to the richness and variability of the subject matter" (Greenhalgh, 1997, p.155).

Introduction

Yin (1994) recommends the final preparation for data collection is the conduct of a pilot study. The purpose of the pilot study or dress rehearsal is to verify the research design by highlighting refinements required of the "data collection plans with respect to both the content of the data and the procedures to be followed" (*ibid.* p.74). The main benefits derived from the pilot study exercise were: assistance in developing relevant lines of questions and conceptual clarification.

The Pilot Case Study

A local business was chosen for the pilot case study. This selection was due to the pilot site being geographically convenient (Yin, 1994, p.74), and hence unrelated to the final selection criteria described in the next section.

After the initial contact was made, but before the interviews, a letter was sent to the Chief Financial Officer (CFO) describing the research purpose, the list of interview questions was attached. A time and date was agreed upon when both the researcher and research supervisor could attend.

The first interview was with the CFO of the organisation. Although the CFO defined the 'model' as a data warehouse, the model was in fact three data marts (or data cubes), which is typical of such New Zealand projects. The company had outsourced the development of the data warehouse. The interview commenced after reviewing the research purpose with the CFO. We discussed the research purpose and mentioned that we were particularly interested in the data modelling stage of the data warehouse development. The CFO remarked that the data modelling had been handled by the company that built the data warehouse. Nevertheless, it was important to hear from her perspective how the activity transpired, as she was heavily involved in the requirements definition stage of the project. For example, the CFO described their involvement,

> "We were involved in what they were doing, obviously, to make them very aware of the idiosyncrasies of the data structure, so they understood how we wanted to see the data in its final form".

Content Modification

It was evident from the pilot study that the participant, who, was also a user of the data mart, did not fully understand some of the terms used in the discussion. For example, terms such as: conceptual data model, physical data model and semantic accuracy[22] were confusing to the participants. Ideally, interview time should not be spent explaining the questions. Consequently, a glossary was mailed or delivered to the participants (attached to the questions) prior to the real case study. This was to ensure that the participants had some understanding of the concepts, such as: enterprise wide data model, conceptual data model, physical data model and semantic accuracy.

Useful questions which emerged naturally from conversation, and were initiated by the research supervisor, were:

1. What kind of questions do you ask of the (sales) cube?
2. Had you read or have you since read anything on data warehousing in any industry - computer journals, for example?
3. How many people have access to the data warehouse?
4. Since you have been running the data warehouse, has it been generally good?
5. Has going through the activity of building the warehouse altered in any way the how the business is run?

Procedural Modification

About half way through the interview a discussion about the background of each data cube was provided by the CFO. It was useful to understand some of the background information relating to the data marts. This type of discussion would be beneficial at the beginning of the interview, or where possible collected before the interview.

[22] These concepts are defined in the glossary at the end of this report.

Therefore, changes were made to the order of interview questions for the 'real' case study.

Conceptual Clarification

Yin 1994, stipulates that undertaking a pilot study may provide "some conceptual clarification for the research design" (p.74). This pilot study certainly did clarify the conceptual framework, and as a result the research question was formalised and a first cut conceptual model was created. The main purpose of reflecting on the research purpose and question is to increase the validity of the research (Greenhalgh, 1997).

As discussed in chapter four the research question and purpose transpired after several iterations. The research question was refined five times due to the pilot study results and insight from further literature analysis. The preliminary data quality framework used to describe the pilot case study findings is the quality framework proposed by Atkins (2000) (refer to chapter two). This framework was used to describe at a high level some of the findings from the pilot case study.

Quality	Quality Goal	Methods
Perceived semantic quality (of the domain).	Semantic completeness	High user **involvement** throughout the project. The users' **understood** the business well, therefore were able to **participate** actively in identifying the project scope, resulting in a preliminary dimensional map that needed minimal changes.
Perceived semantic quality (of the domain).	Semantic correctness	High user **knowledge** of the business domain. (A successful business).
Perceived semantic quality (of the model).	Semantic completeness	Three cubes were developed initially, which provide all the sales forecasting data they need. (Now the users can develop their own cubes). Once the main fact table was created, (identifying all the KPIs), it was easy to add extra dimensional information.
Perceived semantic quality (of the model).	Semantic correctness	The process of creating a preliminary dimension map and then refining that through a prototyping approach, worked well in this situation for ensuring correctness. The final dimension map is correct. The users knew what they were going to get there were no surprises. Users and developers had a good understanding of the current databases (information content of the physical model).
Pragmatic quality	Pragmatic correctness a) understandable b) understood	Understandable as they have been able to change the way they forecast and promote products. Accurate sales forecasting is due to the availability of **accurate**, **detailed data** at the right level, which was understood by the users.
Procedural quality	Procedural correctness	The project scope involved the creation of a preliminary dimension map, which was refined using a **prototyping approach.** The underlying **relational model** (for Impromptu) was developed after the preliminary dimension map was created. For such a small project, the prototyping approach was both appropriate and successful.
Social quality	Social Agreement	**High user involvement**, the sales managers 'owned' the cubes. They were **gently persuaded** to use different methods for seeing and analysing their data.
Structural quality	Flexibility	The data marts have proven to be **easy to change**, the **users participate** in this process.
Structural quality	Simplicity	They are working on developing further cubes themselves.
Syntactic quality	Syntactic correctness	NA (outside the scope).

Table 6: Pilot Case Study Findings

In the pilot case study, both the users' knowledge of the business domain, and their ability to communicate that knowledge to the designers was high. This knowledge had a direct impact on the designer's ability to create a data cube that contained information that was meaningful to the end users. In the quality framework, this is the perceived semantic quality between the participant (business end user) and the domain. This may be represented on the diagram as the distance between the two. For example, the more educated and the more involved the users are, the greater the chances are for success. The designer acknowledged this when he commented, "part of its success is because the users were educated as part of the project", and therefore "that's what contributed to the accuracy of this model, they knew what they were going to get".

The designer discussed the pragmatic approach to data mart validation, he commented,

> "The approach is always to work back from the end user requirements. That often involves going through this prototyping and focusing on an area of business benefit like sales analysis and working out what the KPIs are, and what the dimensional structure is and then working back and creating the mart to support that data".

He also discussed the importance of involving the data consumers in this prototyping process,

> "So we're doing this sort of process with the consumers of the data and not with IT technical designers. We are trying to analyse the requirements with the final consumers, so this process is perfect for that because it gets them involved early - they see immediate results they can't wait to have real data, ...but if the data is slightly wrong, if you've got that the buy-in it's easy to go back and adjust a calculation or get more data. If you've got the buy in - validating this way makes the project more successful".

Although the process for assessing semantic accuracy was not a formal one, it was successful, as reflected by the projects success,

> "It all fell into place for us for this project, because we had everybody there and a model. They put the effort in. In larger places, sometimes there is a delay with key people being away".

The CFO also described the success of the project,

> "I still can't believe how quick it was and how almost instant it was to get those results, it's been one of the most amazing things we've put in. Everybody in the organisation thinks the same thing... It has changed the way the sales people are forecasting how well a promotion is going

> to do. It's also given us from an accounting and administration point of view up to date information, as we do our own accounting at month end for example it gives us up to date information in the way we want it to, so it's just been amazing".

While the dimensional model is simple, it is also flexible,

> "It's very easy to modify the map using the prototyping approach. The only problem we have had that it is so easy to understand and well documented that they have not called us back to do any more work. They can do it all themselves. So they have been developing it and adding more cubes and doing more work themselves".

The purpose of this pilot study was to verify the case study design and to determine criteria for selecting appropriate projects for the detailed study. It was very useful for this purpose.

Formal Case Study Criteria

When determining suitable types of organisations for study, we documented features required of a project. Greenhalgh (1997) too recommends we should "deliberately seek out individuals or groups who fit the bill" (p.157). Miles and Huberman (1994) take this further when they comment,

> "Sampling involves decisions not only about which people to observe or interview, but also about settings, events and social processes. ...Qualitative studies call for continuous refocusing and redrawing of study parameters during fieldwork, but some initial selection still is required. A conceptual framework and research questions can help set the foci and boundaries for sampling decisions" (p. 30).

Sampling in qualitative research is neither probability based nor convenience sampling. Patton, (1990) describes sampling in qualitative research as criterion-based selection. The purpose of this is to conduct an intentional selection process to choose specific settings, persons or events that could not be obtained as well through other methods. To make a purposeful selection, criteria were defined to classify the types of data warehouse projects applicable for study. (However, because of the problems with finding and gaining access to such projects only one case study was undertaken).

Identification of Stakeholders

The unit of analysis for case study research may be an individual, a group, an organisation, or it may be a phenomenon. Giannoccaro *et al.* (1999) undertook a case study investigating stakeholder perceptions of data quality in a data warehouse environment. They interviewed four stakeholder types: data producers, data custodians, data consumers and data managers. Darke *et al.* (1998), also suggest, "The unit of analysis must also provide for sufficient breadth and depth of data to be collected to allow the research question to be adequately answered" (p.280). Therefore, to ensure the breadth and depth of data to be collected is sufficient, the unit of analysis for this project, the stakeholder groups, will include the following: project manager (data manager), data modeller (data custodian), data production analysts responsible for sourcing and transforming data (data producers) and end users (data consumers). This unit of analysis was chosen in accordance with the research goal of determining the importance of semantic quality during data warehouse development, which requires participation from several stakeholders involved in the project development (Giannoccaro *et al.,* 1999). The final set of questions addressed each person's role in the data warehouse project, and many of the questions were elaborated further by a short description (refer to appendices one, two, three and four). This was to capture the different perspectives regarding the semantic quality of the data warehouse.

As one of the goals of this research is to suggest useful strategies for semantic integrity, questions were asked regarding the data warehouse design methodology. In addition, it was important to interview the data modeller(s) to explore the relevance of data modelling to data warehouse design. Furthermore, it was vitally important to interview the data warehouse end users, because they were in the best position to assess how semantically accurate the information in the data warehouse is. This was to determine how effective or otherwise the data extracted from the warehouse is for supporting their day-to-day business requirements. Therefore, the unit of analysis selected for this research coincided with the goal of the research. Likewise, the types of questions asked in the interviews also coincided with the goal of this research by focusing on methodology (process), data model verification and semantic accuracy.

Number of Cases

Originally, the intention was to study a number of organisations (four) to compare projects based on some scorecard of effectiveness. However, due to the difficulties with finding suitable projects willing to participate in this research, we were only able to undertake one detailed case study. However, as it worked out, the time it took to analyse and document the data collected was lengthy, and the main benefit of performing just one case study was that we were able to "investigate phenomena in depth to provide rich description and understanding" (Walsham, 1995).

Walsham (1995) describes a type of generalisation whereby "specific implications in particular domains of action" (p.80) may be drawn through an in-depth case study (for example Walsham and Waema, 1994). While the case study was not an in-depth, longitudinal study, it was a detailed case study. Therefore, like Walsham and Waema (1994) a number of implications (or strategies) are drawn based on the detailed case study.

Nevertheless, it may be an interesting and fruitful area of future research to investigate other organisations according to the following criteria for project selection.

Future Research - Project Criteria

To determine the importance of semantic integrity during data warehouse design, it may be constructive to compare different approaches used by practitioners. It may useful to study and compare projects deemed successful with those whose outcomes have been less successful. However, what constitutes success? Because this research looks at data model validation as a key factor in the 'success' of data warehousing projects, and because such projects are rarely wholly successful or unsuccessful, an appropriate way of classifying such projects for analysis is by some 'scorecard' of effectiveness.

Scorecarding is an appropriate measurement technique in a necessarily subjective area of metrics, and provides a means either, of eliminating candidate projects where there are widely disparate views of effectiveness or complexity; or, of further comment based on the role of the interview subject and his/her assessment of effectiveness.

How do we define whether a project has been effectual or ineffectual? According to the Collins dictionary (Makins, 1995), effectual is defined as: "*adj.* **1.** Capable of or successful in producing an intended result; effective". For the purposes of this research 'effectual' is defined in a data warehousing context as: capable of or successful in meeting the end users' data requirements through providing information that is semantically accurate.

To determine the importance of semantic integrity during data warehouse design the following criteria were decided upon:

- the data warehouse contains data that is: 'meaningful' to the users, correct and unambiguous (Shanks and Corbitt, 1999).
- the data warehouse contains all the necessary data (comprehensive).
- the data can be retrieved in a manner that makes sense to the users (inevitably means the data is stored in a manner that meets the end users needs although they may be shielded from this level of detail).
- the data is captured at the correct level of granularity.
- the end users can produce all the reports/analyses they require.

Effectiveness was further mapped against project complexity. Complexity in this research can be defined in terms of the following criteria:

- ease of access to the underlying source data
- number of disparate databases to be attached to (where the links between the operating system and the data warehouse are complex).
- number of different categories of end users
- size and scope of the dimension map/model
- number of modifications required of the source data

The absolute measure of complexity should be defined after the initial sampling of candidate projects for inclusion. Table 7 below shows a possible classifications scheme for data warehouse projects.

	Simple Project	**Complex Project**
Effectual Project	Effectual + Simple	Effectual + Complex
Ineffectual Project	Ineffectual + Simple	Ineffectual + Complex

Table 7: Four classifications of data warehouse projects.

Use of such a classification scheme may be worthwhile for comparison purposes. For example, a comparison between an ineffectual complex project and an effectual complex project could be made, likewise a comparison of an ineffectual simple project and an effectual simple project. Comparing such a choice of projects may help determine the importance of a validation process for the semantic content of a data warehouse. However, this is an area for future research.

The project investigated for the pilot case study may be classified according to this perspective as simple and effectual. Whereas the data warehouse project investigated for the detailed case study may be classified as complex and ineffectual. However not all aspects of the project were unsuccessful. A detailed analysis of this case study is provided in chapter seven of this report.

Summary

The small pilot case study proved to be a very useful exercise, as it generated changes to both the content and the process of the data collection. Most importantly it provided conceptual clarification, for example the research question and interview questions were refined as a direct result of the pilot case study (refer to chapter four).

Despite the informality of the pilot case study, it did highlight some interesting methods for the success of a simple data mart project. For example to start with the data consumers had a good knowledge of the problem domain, and they were further educated as the project commenced. In addition, a prototyping approach was suitable and appropriate in this situation for verifying the semantic content of the physical data model.

Formal case study criteria were defined, as recommended by Greenhalgh (1997), Miles and Huberman (1994) and Patton (1990). Criteria were specified which directly correspond to the research goal of: determining the importance of semantically integrity during data warehouse development. As a result a classification of different types of data warehouse projects was achieved. However, because of the difficulties experienced with finding and gaining access to such projects, only one project was studied. Several other organisations with current data warehouse implementations underway, were contacted however, they were not suitable either because of the geographical location or because the managers were extremely difficult to access. However, another local project was investigated and one interview was performed. The project was abandoned, as it was another small data mart implementation that may not have provided further implications for this research.

This chapter has provided some of the groundwork for future research by classifying types of data warehouse projects appropriate for this research question. If more studies are performed generalisations leading to prediction may result. However, this research describes generalisations as **tendencies** rather than predictions (Walsham, 1995, p.80).

7 Investigating Semantic Integrity: A detailed case study

When traditional computer databases are used to store knowledge, the conceptual design of the database fixes the semantics and makes it difficult or impossible to reinterpret stored data. This is a problem, for example, if the computer system is used to support strategy processes, business intelligence, or creation of new product designs. In all these cases, information is ambiguous and equivocal - not because we lack information, but because the world is not ready, but under construction" (Tuomi, 1999, p.113).

The Case Study Setting

The case study investigated the design and use of a complex data warehouse at a Government Department. In 1996/1997 a new system, AMS, was developed to record and manage the processing of all applications. The system was custom built using Microsoft Sequel Server 6.5 relational database management system. At the same time the data warehouse was built to provide better access to management information, the rationale was to improve the accuracy of data for business monitoring. The system was to complement the functionality within AMS by providing "management worldwide with timely, accurate and relevant key performance indicators extracted and summarised from AMS" (Lawrence, 1997a, p.1). Before the development of the data warehouse manual methods were used to extract the data, and as a consequence business monitoring was inaccurate (for example, the manual counting of applications was notoriously inaccurate). Therefore, the motivation for the data warehouse was mostly from a senior management perspective to improve business monitoring.

The source database (AMS) is very complicated due to the complexity involved with processing applications and the underlying legislative requirements. The information captured in the AMS system is the source of all information for the data warehouse (MIS). Mair (1999) reported that the data warehouse was developed to meet the needs of three types of user groups: senior managers, analysts in the national office and branch managers. These stakeholder needs are shown at Table 8.

Stakeholders	**Types of Information**
Senior Managers	Summarised indicator information that signals to managers when there are issues to be dealt with.
Analysts in the National Office	Analysts require **useful** information to provide to senior managers, ministers, MPs and members of the public. Moreover, to support research and for producing policy advice to the Government.
Branch Managers	A progress-tracking device for monitoring the numbers of applications processed. This includes details on how long individuals are taking to process applications, how many applications are being processed, the types pf applications that are being submitted and indicating to the manager when the surplus has got so big that they should be sent to another office for processing.

Table 8: Stakeholder Data Requirements (adapted from: Mair, 1999).

The data warehouse was also developed using Microsoft Sequel Server 6.5 relational database management system. Originally, the developers designed and implemented local data marts at each branch office and a consolidated data warehouse at National Office. The intention was that personnel at each location could "access their local datamart using ***Impromptu*** and analyse their summarised key performance indicators via ***PowerPlay*** multi-dimensional PowerCubes" (Lawrence, 1997a, p.1). According to Mair (1999) the cubes were,

> "designed to provide summarised information to Senior Managers, Branch Managers and Analysts. In addition, Analysts who need more detailed information than the PowerCubes provide can use Cognos Irnpromtu, which is an SQL based report writer, to produce their own reports" (p.11).

He also notes that the combination of PowerPlay PowerCubes and Impromptu ad hoc reports were supposed to address the needs of the three user groups. The MIS architecture when it was first developed is shown at Figure 10.

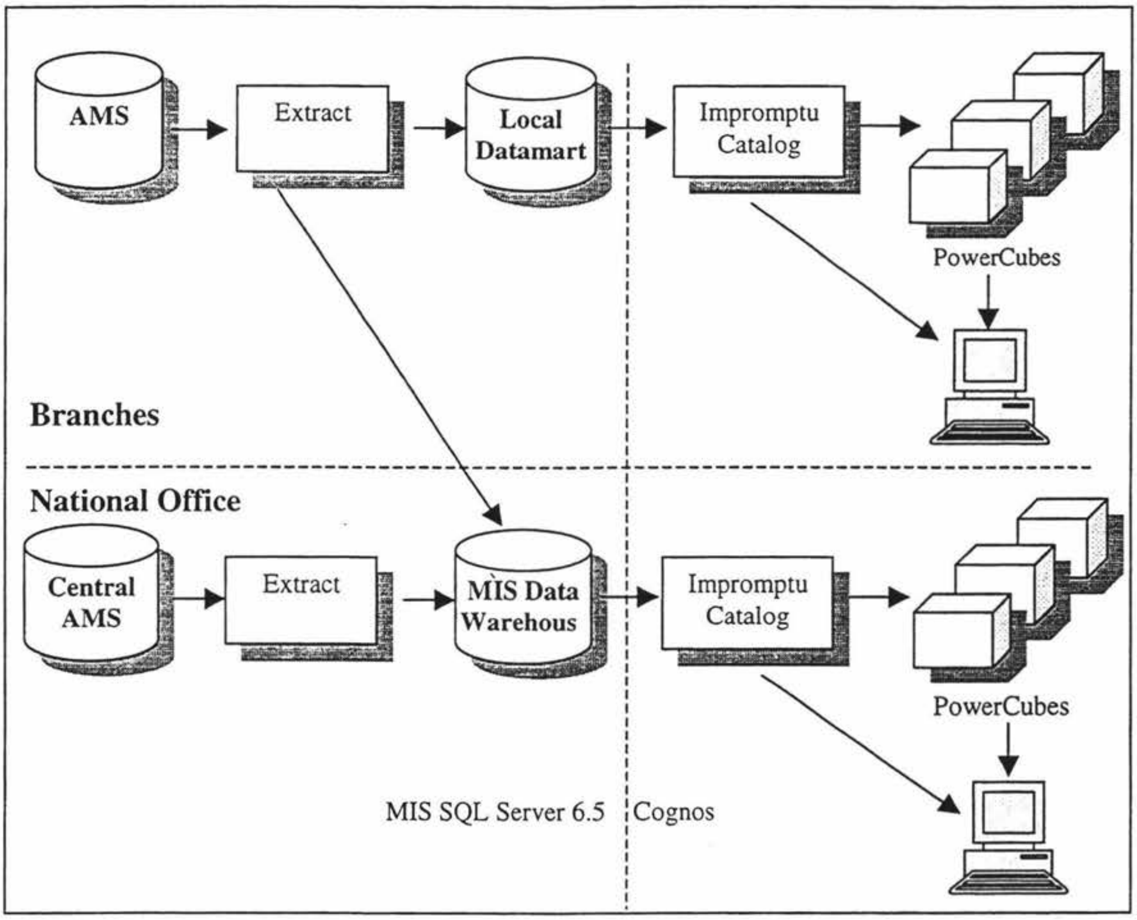


Figure 10: Initial MIS Architecture (Lawrence, 1997a)

The architecture has changed somewhat since then, the cubes are no longer generated locally at the branch level. However, currently there are fifteen PowerCubes held centrally, each addressing a different business need. The current MIS architecture is shown at Figure 11.

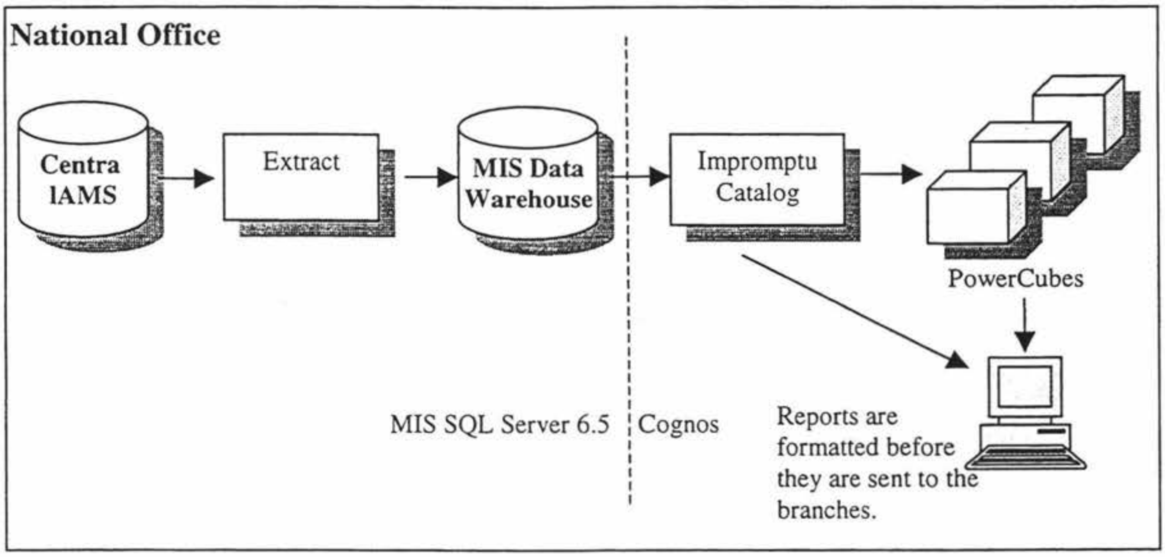


Figure 11: Current MIS Architecture.

The MIS data warehouse was to act as a first level summarisation in the presentation of management information to the business. The data warehouse was also to act as the source for the 'structured' business views (the PowerCubes) and as a source of information for ad-hoc, 'unstructured' query and reporting (supported by Impromptu). The position of the data warehouse between AMS and the PowerPlay business views is shown at Figure 12.

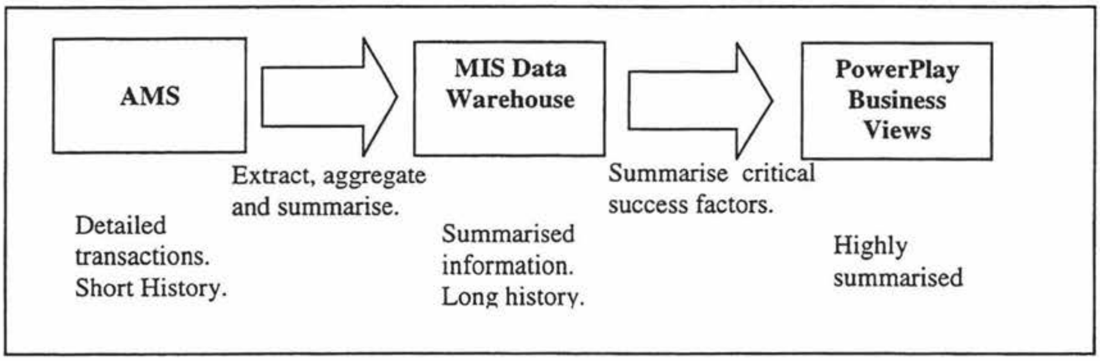

AMS
Extract, aggregate and summarise.
MIS Data Warehouse
Summarise critical success factors.
PowerPlay Business Views
Detailed transactions. Short History.
Summarised information. Long history.
Highly summarised

Figure 12: Data Warehouse Approach (adapted from Lawrence, 1997a).

Painting the Picture

N4 (NUD*IST) a qualitative data analysis tool was used to analyse the data. Firstly an index structure was created based on the interview questions, the relevant sections were coded according to this structure. However, this approach was soon abandoned because it simplified the data too much by structuring it according to themes of questions, versus the more fruitful way of coding according to the themes emerging up from the data. Once a set of free-nodes was established (identifying the key data warehouse use and design issues), the nodes were grouped according to the meaning levels of the framework described in chapter two of this report.

The participants in the case study were: the data warehouse designer[23], a senior consultant, the IT director, a technical support user and four data consumers. It was vital to discuss the data warehouse design and use with the data consumers, the data consumer participants were: from a branch perspective, a research analyst's perspective and a business planning perspective. These covered the user groups identified Table 8 however, due to time constraints and difficulties with gaining

[23] The designer was also the data modeller for this project.

access to the users, only a small sample from each group were able to participate. Although a small branch perspective was gained by interviewing two people at a local branch, a more in-depth analysis would involve interviewing all branch managers at the each of the twelve onshore and ten offshore offices. This was not possible within the resource and time constraints of this research.

A previous case study of the organisation performed in 1999 provides an excellent source of data triangulation, however this work also highlighted some contentious issues.

Generation of Meaning from a data model

Most of the case study participants were not involved in the data modelling activity therefore it was inappropriate to discuss the data model meaning from the users perspective. (Indeed, this is an interesting issue and may be an area for future research[24]). Consequently, the generation of meaning from a data model was determined by discussing the use and meaningfulness of the information (the physical data model) in the data warehouse.

Understanding:

This is the primary meaning of the information in the data warehouse it corresponds to the semantic content of the information. (Refer to chapter two of this report for a detailed explanation of this meaning level).

The organisation has an MIS unit who support the branches by producing reports on a weekly and monthly basis. They believe that as an MIS unit they have a good understanding of the data. However, most other users of the MIS information rely on the unit to access and format that data. For example, branch managers do not directly use the data warehouse, instead they are dependent on the MIS unit to supply the information. The MIS unit run queries and produce spreadsheet reports based on the information in the data warehouse, this is help provide information in a format which is easy for the non-technical users to understand. Participant A commented that, "most of it is really useful, but it is presented in a way which is difficult to comprehend". Although participant B commented that,

[24] An interesting question to investigate might be: how is meaning incorporated in the activity of data modelling?

> "It's that old problem in terms of interpretation and how meaningful it is and where is the connection between two seemingly disparate sets of data, or are they two different things. It is that problem about supermarket warehouses, say on Friday evening the sale of nappies goes up and the sale of beer goes up, does that mean that toddlers are starting to drink beer? It's that kind of stuff and we have a lot of that".

From a small branch perspective participant D said that the reports the MIS department produce at weekly and monthly intervals were easy to understand,

> "The basic data is pretty easy to understand, because it is boiled down to a really basic sort of level. In terms of what we are needing here we've got to process so many residents applications in the year, so it's telling us how well we've done. In terms of that it is pretty easy for us to interpret and to use to determine how well we are going".

So from this point of view the information is understandable however, this is usually after some transformation into an Excel spreadsheet format, so it is not a direct understanding of the data as presented in the data warehouse.

Similarly participant C remarked that she had a good understanding of the reports (based on the data cubes) provided by the MIS department. She commented that the cubes filter out the complexity of the data, which results in a 'face-value' understanding of the information versus having a detailed level of understanding.

From a research perspective participant E thought he had a reasonable understanding of the data in the data cubes, although with regard to the wider context commented that, "it's an extraordinary useful pool of data, but that is balanced against not really knowing what is going on" and later said "but then its just that I don't understand it properly".

A number of factors identified from the interviews that inhibit or prevent an in-depth understanding of semantic content are shown at Table 9. These were classified as either: technical, cultural, training, resource, data or design related.

Factor Type	Inhibiting Factors	Participants Source
Technical	Complexity of the software	A, B, C, D, E
	Teething Problems	D, G
Cultural	Aptitude and motivation	A, B, D, F
Training	Training inadequate	D, F, G
Resource	MIS Resource	A, B
Data	Comprehensiveness	A, C, E
	Format / Granularity	A, B, E, D
Design	Query design	A
	Limiting Reports	A, B, C, E
	Pre-defined cubes	A, B, C, E
	Complexity of the AMS system	D, G
	Use of meaningful codes	E, F, G
	Application typing	G

Table 9:Inhibiting Factors for Understanding

Technical: Complexity of the Software

A recurrent theme throughout this case study is the complexity of the software and consequently the difficulties that the business users experience with understanding the information produced by the MIS.

Participant A discussed how although the latest front-end tools are supposed to be easy to use, they often require the users to have some knowledge of the underlying design.

> "A lot of these front ends, they are not writing code, but they are still a query by example, so people still need to have an idea how tables link together and doing filters and criteria and the average business person can not do that".

Participant A also mentioned that he always got meaningful answers to queries if he persevered long enough, "I just keep re-writing them until I get the right data". However, they cannot expect the average data consumer to have the time or the knowledge to perform such queries, participant A agreed when he commented, "yes, exactly, they wouldn't even understand the concept of relational databases". He believes that even though the data is easy from an MIS perspective to understand and access "somebody in a branch wouldn't know how to use these products, they would find it quite difficult, even though Impromptu makes it quite simple for you".

As it stands, the MIS toolset is no longer used at the small branches to the extent that some branch managers do not know how to access the data warehouse. Participant E commented that the data was not easy to access and understand because "you have to put quite a bit of extra effort into using Impromptu and PowerPlay. Even though I use it daily I still find it quite cumbersome". He also did not know of anyone who actually uses the data warehouse, he suspects this is because "the interface is not particularly intuitive". With a good understanding of SQL, this participant finds,

> "The actual layer of Cognos stuff is probably more confusing and time consuming than just doing SQL queries in AMS. ...I prefer to just write SQL. I find Impromptu really gets in the way, its supposed to be easier but I don't think it is".

Participant E also remarked that he often makes use of other software packages to analyse the MIS data, "I've never actually included raw PowerPlay or Impromptu data in a report". Nonetheless, the MIS data often forms the basis of his research findings. Participant B too notes that the lack of use of the MIS system is partly due to the complexity of the software. He remarks that there is a divergence of views between the developers and the organisation, where the developers think,

> "The people at the branches aren't smart enough to use the product, but really the product is not smart enough to cater to anybody that wants to use it, that's where the problem lies".

He speculates that due to time constraints the business people do not "want to do all these technical, wonderful things" over and above their day-to-day work. Indeed, Mair (1999) also interviewed participant B who in 1999 saw the process of slicing and dicing and drilling down, which are features of the PowerPlay software, as too difficult for the majority of users. In the recent joint discussion session, both participants A and B concluded that the tool set was too difficult for the average business users technical knowledge.

Likewise participant C noted that the data was difficult to access directly because she does not have the software knowledge. However, she does not choose to use the software herself because the MIS unit can access and understand the data on her behalf. She also talked of the frustrations some branch managers were experiencing due to the lack of access to the operational data. She commented that,

> "It is not because the information is not there, we do transport the cubes to them so they can play with the cubes themselves, but they simply don't have the time to invest in learning that software, to be able to pull it and use the information".

However, most importantly participant D comments that the tool set "is reasonably user friendly if you just want to use it as a very simple tool". He also says that although the software supports very complex analysis, ~~but~~ that capability was not relevant at the branch level, "obviously it is a lot more powerful and you can do all sorts of weird and wonderful things, we probably don't need that sort of capacity - not at the branch level".

Technical: 'Teething' problems

Unfortunately, the MIS data warehouse did not get off to a good start, and in fact the cubes that are currently in use are completely different to those that were designed and built four years ago. However, the troubles that plagued the data warehouse initially have been detrimental to the reputation of the system.

For example, participant D claims "there were slight differences with the figures that came out of MIS versus what we thought we had done, so there was some teething problems", however an employee[25] worked well to fix those errors and at the time was an excellent support person. Participant G also said that initially in 1997,

> "problems were with the branches getting out of line with each other and the central MIS, where the central MIS would not give the same answer as the local data mart ...that is multiple data marts coming from different source systems - it didn't work very well, the data got out of sync".

Cultural: Aptitude and Motivation

Participant D has highlighted a rather contentious issue that from a small branch perspective there is no real need for access to the raw data on MIS. The reports provided to the branch support their current needs, so in his opinion they do not need further detail. Therefore, it may be unfair to say that the users do not have the aptitude or motivation to use the tools. More specifically participant D says,

> "For a small office, where you see everybody all day every day you get a feel for what is happening on an individual basis. I'm not sure how the bigger branches use it, they may need to rely on some of the more detailed information on AMS, but we don't necessarily need to have that. I think the weekly reports don't give a need for it".

He continues, "we just need to know what our numbers are that we are turning out and who is turning it out and what percentage is pass and fail. I think the weekly reports can suffice for most of what we need". This view contrasts with Mair's (1999) findings where he interviewed an external party managing the project who "**speculated**[26] that the reason why branch managers did not use the information was

[25] This employee has since left the organisation.

[26] Emphasis added.

because they were uncomfortable with a business model that the MIS implied, because it might make their work more stressful" (p.11). Participant D did not believe this was an issue at the small branch level. Despite this participant B sees the lack of motivation to use the MIS system at the branches as a measure of the systems success. He comments that, "the content is not compelling enough and the delivery vehicle is not easy enough, so either way they are not motivated to go and want it and look for it, so it's has not been successful".

Whereas participant A realises that "some of these people are still struggling with the concept of the computers, it's all about providing them with something they can use, and not expecting them to up-skill". He empathises with the branch managers needs and remarks that,

> "In terms of the tool I don't think you should be giving people like branch managers that kind of tool, they just won't use it. You've got to have low expectations and have empathy with these people when they say that they don't understand it, because I think a person needs a certain aptitude to be pulling information like that. ...There's a whole lot of issues, its about the product we've given them, its the expectation that they should do stuff themselves".

Inadequate Training

Another inhibiting factor for understanding is the apparent lack in training for the MIS system. Regular training should be provided, versus having one off technical courses on how to use the tools. If the end users are to realise the potential of the data warehouse a lot of investment should be made in ongoing training.

Participant F said training was both an issue at the AMS level and the MIS level, and that there was a need for "far better training on AMS". He also said that there was a general lack of documentation or reference material, and that the end users rely on the person who has been around the longest, or has used it the most, to provide support.

From a small branch perspective participant D did not know how to access the MIS, and no longer had the software on his computer, he was "not sure how that came about, that's obviously a training issue"(now they can't access MIS at the small branch level even if they were interested in using the data warehouse). He said that with the advent of AMS "there wasn't the time and there wasn't a lot of training done on PowerPlay itself when it first came in which didn't help, there was some basic instructions to have a 'play' to get your basic data, and that was kept at that fairly basic level". Moreover, participant G recognises the lack of training at the branch level when he comments,

> "The other major reason was that the users never got trained at the branches really, because they were busy trying to cope with AMS (which was hard), that they didn't have time to do anything with MIS".

Nevertheless, Mair's (1999) report speaks of an email survey conducted in June 1999 that found amongst other things:

> "The support, coaching and training provided to branch managers on the MIS system had been largely unsuccessful in increasing the skill level of would be users"(p.11). ...In spite of the fact that many people had received training, they have not used the software and do not feel **confident** in trying to use it now" (p.12).

MIS Resource

Both participants A and B agreed that there had always been a problem with the limited MIS resource. Participant A exclaimed that the problem is not with the data but with the lack of people resource. He commented,

> "We really haven't had information people working closely with the business to provide them with the information they need, there's only 2 of us now, and there was only one. It's also a management issue, nobody has really been managing that process of getting management information to the business and all those issues of how do we provide it".

Data: Comprehensiveness

Participant A remarked that he is not always able to produce the reports and run the analyses he requires, due to how the data is presented. However, he notes that "MIS includes most of the necessary data, but I don't think the cubes have all the necessary data, but I don't think that's there purpose, it should be if the branches are using them, but they're not".

Likewise participant E says that a lot of the data on AMS is not carried onto MIS,

> "There are occasions where you want to know ...for example, what are the demographic characteristics of people approved under general skills category? You can't actually get that information on MIS ...When you're looking at the individual record of data it's quite often AMS that you have to go to after you identify through MIS what it is you want".

However, he also concedes that MIS was not originally designed with research requirements in mind, therefore a lot of the information he would like is not represented on MIS. Nevertheless, participant E says that MIS is not comprehensive as it "would need to be people based ...as opposed to applications based therefore, having everything about individuals easily accessible".

However, participant C finds the data comprehensive (but says it is perhaps not for others) because,

> "I'm only concerned at a summary level, but certainly the complexity of the information it can pull is sometimes limited because it is meant to be a reporting tool, it is not meant to be telling you what tattoos – you know that sort of thing. But over time with operations a lot of managers do actually want that sort of information, it can be important marketing information to have lots of detail and try and profile it somehow".

She specifically noted that the information available for the investigation side is quite limited. However, she remarks that the data is "meaningful on an aggregate level but trying to drill down underneath ...you loose something, it can't capture everything".

Data: Format and granularity

This factor refers both to the presentation of the data and the granularity of the data. The participants seem to agree that, all the data is there, and is reasonably accurate, but it is difficult to use because of the format it is presented in. This could be a result of the software and/or the design. Participant A believes that the data is presented in a format useful for his needs because he had access to very low level data and high level groupings of data, but the problem is providing this information to the business people. Moreover, participant B commented that not only was the data not compelling enough, but that "the delivery vehicle is not easy enough", so the format of the data as presented in PowerPlay and Impromptu makes it difficult for the business users understand. He continued "most of it is really useful, but it is presented in a way which is difficult to comprehend, which is where we need to make it a lot easier, at the front end of what they see. At the moment all the complexity of the back end is exposed to them, and they don't need to be burdened with all that".

Similarly, Mair (1999) reported that participant B "regards the information that MIS provides as data rather than information" and that "the information is too detailed for most users and is therefore not delivering the information that people want" (p.11).

Participant D said "the MIS data is pretty raw so it doesn't give a lot of that background". Nevertheless he remarks,

> "It is at the level of detail we need in terms of knowing how well we are performing in terms of meeting our targets, whether we are likely to meet it or not going to meet it. So it gives us the raw data in terms of that, it is certainly most useful for that. It does not tell us why, why we are not meeting or anything else, those are different issues, but in terms of signalling whether we are going to hit the number or pass the number or be short it is useful".

This is an interesting point and highlights an issue that participant F refers to, "the data should be analysed more to identify trends and patterns, ...this was not being done because of resource constraints" (Mair, 1999, p.11).

Design: Limited reports

One of the major benefits of a data warehouse should be the ability to rely on the production of reports for trend analysis and business monitoring, especially reports that present the data in a format relevant to a particular business users needs. Clearly, there was a problem with reporting when the MIS system was first set up, so now the users rely on the MIS unit to create their reports for them.

Participant A initially thought there were problems with the data in MIS, but he has since discovered it was the way the queries were being written and subsequently the reports given to the business people were wrong. As mentioned earlier participant E (who regularly uses PowerPlay and Impromptu), notes that he always manipulates the raw data further by analysing it in either Microsoft Access, or Microsoft Excel to present the information in a different way.

Participant B concludes that traffic light reporting might be more suitable, for example,

> "The general manager who has 25 branches, he doesn't want to know every Monday every branch what it's doing, (sic) but lets say there is a problem in Singapore he just wants that to be highlighted, so that's the kind of thing that wasn't there in the design".

Similarly participant C also notes that (to a certain extent), the reports that have been set up for general consumption, are limiting.

Design: Predefined Data Cubes

PowerPlay extracts and organises the data from MIS into multidimensional cubes. However, the participants have identified issues with the data cubes implemented. For example, participant A discussed the limitations of the data cubes,

> "It is the age old problem, where a cube has been built and there is some information that you would like to see in there that is not, a cube is really just a giant query that somebody has written, so it depends on how they have written it, some cubes aren't useful, some are. In some areas we have too many cubes doing the same thing using a slightly different methodology".

Participant E also commented that, "you can find out a lot about applications, how many applications say for last month, but it is not so easy to find out how many people we had with work permits. ...because it is hard to construct the query". Participant C alluded to problems with the type of information provided by the predefined data cubes,

> "Because of the way the cubes are structured it is quite pre-defined in the sort of information you can get, and apart from the MIS analysts people don't generally use Impromptu to go in and get the information themselves, because it is so complex".

Design: Complexity of the AMS System (source database)

Another inhibiting factor has been the complexity of the source database AMS. From a management perspective Mair (1999) found that one of the impediments preventing the organisation from changing from its current state to the desired state, was the complexity of the system. More specifically he writes,

> "There is a view amongst some ...staff that even though the ...system has been designed with the best intentions, it is still too complicated and time consuming" (p.6).

Likewise the participants suggested that one of the reasons for not using the MIS system was the impact of AMS (specifically the amount of knowledge required to use the system effectively). For example, participant G remarked that AMS is:

> "A very distributed database, and it's quite a complicated system there's a lot of tables in there. It needs to be very closely tuned, and is designed to work satisfactorily across the wide area network (world wide), so it's not brilliantly designed for reporting".

Design: Structured codes

A major complexity in the AMS system is the notion of grounds codes, which are logically simple, but at the physical level have been designed in such a way that it makes it difficult for the users to select the appropriate data. The use of structured codes[27] in database design often causes flexibility problems and is generally not recommended.

For example, in AMS (and MIS) a grounds code comprises three parts: application category, application type and application criteria. The application category code identifies whether it is a visa, a permit, an appeal or a border application. The next level identifies the application type code (a type might be visitor, work, residence), below that at the lowest level is the application criteria code (for example, general skills, family, humanitarian, family marriage). This structure requires (of the users) an understanding of these codes, so the correct records are selected when performing queries. Participant F said, "logically it is very simple, what follows however between the paper and the system they don't often match". In addition, participant G commented that there were problems with the users understanding the complexity of the grounds code structure,

> "This is a good database design comment ...they had this thing called grounds code, it was a meaningful code ...you can drill either way from these, they've got thousands of these, these guys just wanted to know how many applications they had dealt with, it might be this particular type, plus this particular type, plus this particular type to get the number, so what they needed was a business analyst ...who knew exactly what we were counting at any particular point in time, he had to do some pretty complicated PowerPlay reports to give these people their simple number".

Design: Application typing problem

Another problem stemming from database design issues at the AMS level is the fact that every document is handled as an application even where they quite obviously record different information. Participant G spoke of this problem from a design point of view,

> "The data is an issue too, which is only now beginning to be sorted out. In that, its the way they handle everything as an application, but different types of applications are very different like a visitors visa application goes through a very different process and has different bits of information to a permit issued over the border, or a ministerial appeal".

? defined in next paragraph.

[27] A better technique may be to store these three codes separately, (or if they are going to merge them they should not expect the users to know how to drill-down to get the correct numbers).

However, the data warehouse does not need to reflect the design of the source database, the designer now concedes that he would do things differently, including denormalising the data structures more.

Query design

As mentioned earlier there have also been issues with the Powercube query designs. For example, participant A initially thought that the data was inaccurate, but has since discovered that "it was the way that queries were being written". These types of problems have been identified and worked on, although participant A said, "somebody that receives reports from us or has to do their own reports from the cubes might say the data is inaccurate, the data may very well be inaccurate, but it may be a query design".

Validity Claim - Comprehensibility

Habermas' (1984) validity claim for understanding is comprehensibility, following Mingers (1995). Participants A, B, C and E all made references to problems they or the business users they support, have with comprehending the detail of the data. If comprehensibility is a problem, then "it may reflect a lack of *structural coupling* (Maturana, 1978) between the" designer and the data consumer "- the signs do not have common connotations" (Mingers, 1995, p.299).

Strategies for assisting understanding based on these inhibiting factors are presented in chapter eight of this report.

Connotation

This meaning level examines the complex of other meanings, beliefs and implications associated with the primary meaning of the data in the data warehouse. In the context of this research it was not possible to directly measure connotation with respect to the data model, but rather the information provided by the data warehouse.

Differentiated groups of users

Shanks and Corbitt (1999) point out "the social construction of the meaning within a society is determined by the cultural practices of that society" (p.791). Mingers (1995) talks of knowledge and experience of differentiated groups of people, and how they share common understandings unavailable to outsiders. This organisation certainly employs a diverse range of people, some of whom require information from the data warehouse. Participant B talked about the community of users at the branch level,

> "I don't think the cubes are necessarily the best solution for people out in the branches, which would the majority of our user's, our branches at the moment are keeping their own manual logs, because – what does that mean? It means that MIS has failed to deliver that information".

Participant F also talked of the different reporting needs of the user groups. Specifically, he discussed the reporting needs missed at a high level (reflecting what is set out in the annual report). He also discussed the reporting needs at a lower level, where branch staff are interested in completion numbers on a day-to-day basis (also, how many are still on hand), but then at a high level, the senior managers are not interested in that type of detail. Furthermore, he mentioned the needs of the policy people in the National office, who are interested in the same type of data but from a different **perspective**,

> "They want to look at not just numbers, but the break down of those numbers, in some respects they are looking more at personal-bio side rather than just counts, which is what the purchase agreement was".

Therefore, the business users of the organisation have very different reporting needs, although their needs are based on the same data. Originally, they tried to meet these differing needs with the same set of reports. Participant F said that it had taken them four years to even come close to meeting these different requirements "it's taken them a long time to get to where what really should have been available, or was available, to getting it understood from day one". In terms of the data warehouse the designers can guarantee that there is the same number of applications in MIS as there is on AMS, but whether that is the correct number that the users want to see is the real issue.

A number of inhibiting features for connotation are classified at Table 10.

Factor type	Inhibiting Factors	Participants Source
Social	Access to data (Political)	D, F
	Capturing different user needs	E
	Inconsistencies (Cultural)	F
	The wider context (Cultural)	D, F

Table 10: Inhibiting factors for Connotation

Mair (1999) discussed cultural issues in his case study, he wrote,

> "The implication is that the present problem is not the information, but cultural issues alluded to previously. In addition, some people consider

> that user needs are not generally met ...because users are not consulted adequately about what they need and want. Instead, senior managers attempt to mandate what users will have, causing user protest by boycotting new systems and methods" (p.13).

Social factors: Access to the data (Political) → is this connotation ?

Participant G said that initially there were problems with the data cubes at the branch level, "originally they wanted all the information, to know how all the other branches were doing, that got into politics as how much data they were supposed to see, they were competing with each other". However, participant D said that competition did not exist between branches now,

> "Historically, they have tended to see how like branches are doing, to see how well they were doing, but the data is pretty raw, it wasn't used to say that one branch was better than another. There were some charts that tried to rate these sort of things, but because it was based on raw data it didn't really take into effect other complexities which may have had those sort of effects".

Social factors: Capturing different user needs ✓

Because MIS was originally developed as a business monitoring system, some problems have come about due to the changing use of the system. Although, at the time of development it may have been difficult to predict the future use of the data warehouse, it may have been useful from a design perspective to have utilised an approach that allowed for flexibility.

Participant E realises there is an issue with the original purpose of MIS and meeting his research needs when he comments,

> "I am using it for purposes other than it's original purpose, there is a little bit of dissidence going on. For example, I think MIS was designed to be more of a business reporting type of thing, ...monitoring applications and working out how timely things were. I'm using it for a completely different use, I'm asking policy questions ...for example, I'm probably more interested in looking at individual clients however, MIS is really designed to look at applications so it's that traditional thing of using administrative data for research purposes".

Social factors: Inconsistencies ✓

Participant F commented that not only was the length of time to process applications inconsistent between branches, but also the quality of application processing was inconsistent between offices too,

> "You start looking at the quality of applications, if you look at nationality by application type, why would it differ between branches, the approval and decline rates? I was reporting on that to a policy and national level and sending it out, but it was too much for the branches they didn't want to know that, if it was me I would like to make sure".

He continued,

> "Then you've got a greater decline by some branches off shore so they have a specific nationality or ethnic group going through there, but when you look at the same nationality group within New Zealand offices there was a higher approval rate off-shore than on-shore, so why was that? Did they know other things off shore that they didn't know on shore?"

If such a use for the data warehouse was possible and demonstrated to the users, they may feel more inclined to find out such information. However, staff at the small branches may never use the tools if they do not need the information on a day-to-day basis for processing applications. Currently the reports provided suffice, they have no other current need. Nevertheless, their needs might change and subsequently the use of the data warehouse at the branch level may increase. In defence of these claims participant D remarked that no two branches deal with exactly the same market, making comparisons difficult,

> "Even though we are all operating the same sort of policies, there is ...subtle differences in terms of the markets we deal with. If you are dealing with a high quality person coming in ...then we are obviously going to churn out a lot more, whereas if you're getting applications under a different policy like our humanitarian category it takes a lot of in-depth work, the numbers that you churn out can be somewhat less. It then becomes very difficult to compare numbers against numbers."

Social factors: The Wider Context

Participant F asserts that the staff at the branch level lack knowledge of the wider context. Even so a participant[28] from a small branch explained why he did not use MIS,

> "I don't use it on a day to day basis, it's a managers tool, its not really useful for our job, we can process an application fine without MIS, MIS is good for trends and recording numbers and things like that, when I'm processing an application I don't need to know what is in MIS".

[28] The discussion with this person was not coded because he had never used the data warehouse, nonetheless it was interesting establishing why this was so.

Participant F exclaimed that the quality of staff employed by the organisation differed by region, and that if the staff "were really interested in reporting and creating more value from their data there is a lot more they could do".

Validity Claims

According to Mingers (1995) this level concerns Habermas' (1984) validity claims of truth and rightness. This involves determining whether the "propositional content of the sign is actually **correct**", the "state of affairs actually exists" and whether "its claims about social rules and roles **acceptable**" (*ibid.*p.300).

According to participants A and F, the organisation has had issues with data correctness because people have sometimes not entered data in correctly at the AMS level. Although these types of problems are outside the scope of the data warehouse design, it is important to recognise them as issues and possibly to provide suggestions on how to minimise these (for example, domain restrictions, preventing a user from entering a record twice on AMS, to name a few).

Some managers at the branch say the data is incorrect because it is not current. Indeed participant B thinks the fact that the data is updated on MIS weekly is a flaw, and they should be provided with all the information they want for operational purposes more immediately rather than waiting for a week. Likewise participant C commented that "managers are managing their stock and applications on an hourly basis but they don't have the information to match, so there is a level of frustration there". She said the other issue the organisation has with the data is, "a back loading issue - the time delay it takes for a branch to enter the information into our system. So it means that at a point of time you've never really got a clear idea of how much is to be decided, or how much work we've got, so that's a bit of a limitation". (However, the original design document specified that the MIS data warehouse was to be updated on a weekly basis).

Despite these issues most of the participants thought that the data in MIS is fairly unambiguous. For example, participant C said, "I think, the business at least has a fairly set understanding of what we mean when we say a decision or an application received, there are levels of jargon around, I would say it was pretty unambiguous".

Intention

According to Mingers (1995) this is the individual, subjective meaning for a particular person and the implications of that meaning for action. These are the intentions it will lead the person to have, "personal experiences, feelings and motivations at a particular time will be brought in and result in a particular activity" (*ibid.* p.300).

The organisation needs to realise that from a small branch perspective there is no current motivation to use the MIS and Cognos toolset because the weekly and monthly reports provided suffice. Two participants noted that from a small office perspective they have 'hands-on' knowledge of staff performance and application numbers, without needing to use the data warehouse. These reasons may be valid, and unless they are required to use the MIS system for some aspect that supports their day-to-day work, then the use of the MIS system at the small branch level will not change. Nevertheless, providing better training to identify how the tool set might assist with their day-to-day operating may prove fruitful.

The inhibiting factors for the meaning level intention are shown at Table 11. However, all the previous inhibiting factors for understanding and connotation may also be inhibiting factors for intention.

Factor Type	**Inhibiting Factor**	**Participant Source**
Design	Range of reporting requirements	G
	Reports not compelling	B
Technical	Response Time	A, B, C, D, E, G
Cultural	Reliance on one person	A, B, C, D, F,

Table 11: Inhibiting Factors for Intention

Design: Range of reporting requirements

From a design perspective a difficulty has been supporting the range of reporting requirements. They need to recognise the importance of identifying the different reporting requirements, and consider the impact this might have on the data modelling activity. Participant G notes, "their reporting requirements varied very differently according to the application type" for example, application types are counted using a specific method over a specific date range. This type of complexity is too difficult for the average business person to understand.

Design: Reports not compelling

According to participants A and B, another inhibiting factor are the reports themselves. They claim the reports are not compelling enough as they are too detailed for most users. More specifically participant B remarked,

> "It seems to kind of produce a lot of reports, and the same kind of reports, and they contain a lot of information, ...it is of interest to some people, parts of it, but its not compelling enough to make people want to use MIS all the time".

Cultural: Reliance on one person

Previously an MIS analyst who understood both the data, and the tools, was relied upon for fixing problems with the design. This person was also responsible for providing ad hoc reports to the business users. This is an inhibiting factor because this key person has since left. Although this person has been replaced the new person has to catch up. But, more importantly the attitude of the end users is to ask for the information instead of retrieving it themselves, the motivation is not there because of problems with the complexity of the data and the tools, therefore the intention is not there either. Participants A and B discussed this issue,

> "We used to have an MIS analyst who found it very useful and everybody relied on him to give them the data, therefore it became more and more complex, he knew how to find his way about, he created his own queries and all that, but there wasn't enough engagement with the people who run the business from a day to day basis at the higher levels".

Mair (1999) also discussed this and found that many managers rely on the reports generated by the analyst in lieu of accessing MIS information directly. In addition, branch users have not had the motivation, or time, to generate their own reports. For example, participant D comments, "I don't have access to PowerPlay anymore. It's basically got to the stage that the data is all evaluated for us, ...they are producing the reports for us monthly rather than us producing our own ones".

Technical: Response time

→ is this at the intention level?

The response time is generally very slow (all the participants mentioned this as an issue). Although this is outside the control of the developers, it may have exasperated other problems. Participant E described the problems with response time,

> "I still find it quite cumbersome particularly the time it takes to run. ...You'll have some query run for twenty minutes and at the end you've got to cross product or something like that, so that whole twenty minute process and you're back at the beginning, that sort of lack of feedback as you go means that it feels quite inefficient".

(Apparently, this problem is currently being addressed).

Validity claims

Mingers (1995) regards sincerity as the primary validity claim at this level. In terms of the data warehouse, this is the reliability of both the AMS and MIS systems.

Participant D commented about initial problems with MIS, "there was this apprehension about the reliability of the data". He claims that MIS did not get off to the best start, nevertheless they still used it "because we didn't have a lot else, so we had to rely on it". Similarly participant G said that initially the users did not trust the data, so the data cubes fell into disuse. Participant A reflected on the fact that the majority of the users at the branches complain about accuracy, and are keeping their own manual records, "they are typing data into AMS and they are keeping their own manual files at the same time". Also participant C said that some users were not entirely trusting of the data. Participant D too said that they maintain manual records to validate the information in AMS.

While participant A said the reliability of MIS was not too bad, there are some problems. For example in terms of data collection, there may be tension at the front office to process the application as quick as possible, "whereas from a back office point of view you want to collect as much data as possible, you have to strike a balance somewhere". Likewise, participant B said the data quality of MIS was "average to good, it's good, but there are quality and accuracy issues in the data, I don't think that has been managed correctly".

From a planning perspective participant C said they have a few issues with the reliability of MIS because of the way it extracts the information from AMS. More specifically, she said "MIS does a weekend extract of the information from the AMS database. ...It doesn't refresh completely, it only refreshes records where it's had a date flag, that, during the course of the week, something has changed with that record ...but in that very process inaccuracies do creep in".

These problems are addressed periodically where the entire MIS database is refreshed, this is so that the data in MIS matches the data in AMS.

Producing a data model from meaning

This is the obverse of meaning generation, that is, data model production (Mingers, 1995). Like meaning generation, three stages are described, from the intention of the data modeller through to the actual creation of the data model. A detailed description of these stages may be found in chapter two of this report.

Intention

This concerns the intention of the sender or producer of the data model. Mingers (1995) says that questions at this stage concern the nature of the intention, and its sincerity. This corresponds to some degree to the user requirements stage of the data warehouse project. This is clearly where the intention should be addressed and where agreement should be obtained through stakeholder involvement.

The original intention of the data modellers is unclear because the MIS database was created in 1997 and there is no direct reference made to their intention in the documentation. However, the data models were to: describe the logical and physical information required to support MIS. Furthermore, the central MIS data warehouse was to hold information sourced from both the Central AMS database and the branches, and to hold a consolidated and summarised history both of work activity at the branches and of applications made to the organisation. (Lawrence, 1997a). The intention now from an MIS perspective is getting the data "out of the data warehouse and into the hands of the people that need to make decisions with it".

The inhibiting factors for Intention are both classified as problems relating to the user requirements definition. The factors are: problems with understanding the business domain and the technical focus of the project. (There may well have been other inhibiting factors for intention).

Understanding the Business domain

In addition, participant F raised the issue of the users not understanding the application process fully, "You just have to look at the rate of appeal, ...they don't know how to process applications or how to apply the policies to an application". Apparently, it takes a long time to acquire the knowledge to successfully process an and

application, this affects both external an internal staff. If the employees do not fully understand the application process, they may have difficulties effectively communicating the requirements to the data modeller.

Technical focus

thought

It is evident from the participants comments that the intention for the data warehouse was from a technical perspective rather than from a business perspective. Participant G said, "they had some rather poorly though out potential cube designs. We got involved because they knew of our products, they took client PowerPlay and Impromptu to fill those roles". This statement reflects the technical focus of the project, where the product potential may have been the focus. Likewise participant B said there was insufficient involvement of the users at the various levels,

> "I think a lot of the earlier design and development was probably driven more from a technical perspective rather than from a business perspective, so there wasn't enough engagement with the stakeholders such as the general manager or the management level that would actually have benefited from the information".

Generation

This stage involves converting the intention into a representation of the data; this is the data modelling activity. Firstly, a relational data model was created based on the tables in the implemented AMS system. Secondly, a star-schema bus architecture was developed as the physical design for the data warehouse. The MIS design involved two consultants from an external company who also implemented the Cognos toolset.

The main data warehouse designer/developer, a senior consultant, had attended various data modelling courses throughout his career and had ten years experience in business intelligence type projects. Prior to the MIS project he had completed designs for both transaction systems and OLAP systems, however the MIS design was heavily influenced by the Kimball (1996) approach[29].

[29] This is largely because his experience had included working with the first OLAP tool Metaphor, which is apparently, where Ralph Kimball started. Interestingly, participant G said that the star schema approach was invented because the tool Metaphor needed it.

A number of inhibiting factors contributing to issues with the design are classified at Table 12 as either design or project management related.

Factor Type	Inhibiting Factor	Participant Source
Design	Analysis specification outdated	G
	Lack of user involvement	A, B, D, G
	Design data model based on AMS	G
	Constrained by the design of AMS	G
	Complication of the model (factless fact tables)	G
Project Management	Project timing	D, F, G

Table 12: Inhibiting Factors for Generation

Design: Analysis specification outdated

As mentioned earlier the project faced difficulties from the start. From a design perspective there were problems, for example the original data warehouse design was based on an outdated requirements specification. Participant G commented that, "we designed MIS initially based on a specification prepared by an ...analyst who had himself designed probably a couple of years before it all came out". Not only was the analysis specification outdated, but also it may not have been what the users wanted and probably "came out of his head". In hindsight participant G said he would not rely on this type of specification again.

Design: Lack of user involvement

Throughout the business requirements definition participant D said he was only involved at a high level, "I had off and on some small involvement, I was part of the steering group, but that was more at the higher level. That was probably my biggest involvement". Likewise, participants A and B agreed that there "wasn't enough engagement with the stakeholders". Specifically, participant B remarked that the development should have been targeted more towards the people who would value the information (the executive level and the branches), and there should have been more interaction with the business. Mair (1999) also discussed the lack of consultation with the business people to determine their needs and wants. He wrote that the users boycotted new systems and methods because the senior managers determined their needs for them.

Design data model based on AMS

The project documentation presents the database logical content and the database physical design, a logical data model is also mentioned (but this model can no longer be located). According to participant G the 'logical' data model was based on a one to one mapping of the source database AMS to the data warehouse MIS. However, according to Date (2000), the logical data warehouse model should be developed from a logical perspective and therefore without regard to the physical constraints of the source database AMS. Upon examination of the documentation, the logical database description talks of "attributes that are required to support the PowerPlay business views" (Lawrence, 1997b, p.5), summary attributes that "relate to those required by PowerPlay" (*ibid.*p.18) and derived attributes. These types of attributes are often described at the physical design stage. In addition to this, entities that "contain weekly snapshots of summarised details" (*ibid.* p.23) are physical design issues.

Furthermore, the 'grounds_code_id' is a system generated surrogate key that uniquely identifies a grounds code record, it corresponds to the: application_category_code, application_type_code and application_criteria_code. Usually at the logical level these three codes would be shown as attributes of the application entity linking to individual lookup tables. Simsion (1994) says the "main two arguments against surrogate keys are programming complexity and performance" (p.213), but more importantly that when specifying surrogate keys that simply specifying grounds_code_id as the key to the groundscode entity does not solve the real world problem of matching real world applications with rows in the database table. Although the grounds_code_id is a surrogate key, it relates to three different attributes on the grounds code entity, the users need to have an understanding of this structure so they can choose the correct category, type and criteria codes when running queries.

Usually the logical structure of the data is how the data is in the real world, but the generalisation of the application entity at the logical level may not have captured the different application types adequately. It is very important to propose other candidate data models, hence it may have been useful to sub type applications at the logical level (even if these designs are rejected at a late stage). Simsion (1994) writes,

> "Our choice of level of generalization will have a profound effect not only on the database but on the design of the total system. The most obvious effect of generalization is to reduce the number of entities, and, on the face of it simplify the model. Sometimes this will translate into a significant reduction in system complexity, through consolidating

> common program logic. In other cases, the increase in program complexity from combining the logic needed to handle quite different subtypes outweighs the gains" (p.86).

However, in this example, it is difficult to determine the pros and cons of generalisation because there is no evidence of more than one workable solution.

The physical design of AMS

The physical design of MIS consists of a number of star schema's (Lawrence, 1997b, p.27). Date (2000) has described problems with using the star schema approach to data warehouse development (refer to chapter three for a summary of these problems). One of the main problems with this approach is the general lack of logical modelling and adherence to good database design practice (refer Date, 2000, pp.694-725). Participant G recognises that the data collected operationally was also a constraint. Apparently, the data held in MIS is limited by how the data is captured in AMS.

Design: Complication of the data ('factless' fact tables)

One complication with this modelling situation is what Kimball (1998) describes as 'factless' fact tables. This is where there are 'no facts' in the fact table because the model is either recording an event or coverage. Participant G remarked that,

> "most of their Powerplay KPIs are counts or averages, like how many applications are there of this type that were finished on this date, handled by these people, there is no revenue for an application, so all the models count things - so we're counting applications".

Although Kimball (1998) suggests that a "fact table consisting only of keys is a perfectly good fact table" nevertheless the SQL for these queries may be very difficult to read. For example, participant E has difficulty with constructing queries using Impromptu, "it is not so easy to find out how many people we had with work permits. ...because it is hard to construct the query". He continued,

> "If you look at the actual SQL behind an Impromptu report it is incredibly confusing, it's like no SQL I've ever seen before. I just feel that a SELECT WHERE would be ...a lot easier than going through all the Cognos content".

Project management: Project timing

The project did not receive the necessary attention from the organisation, as at the time they were preoccupied with the development of the source database AMS.

Although the MIS that resides today is quite different from the one developed in 1997, the attitude towards it perhaps has not changed. Upon reflection, participant G thinks that they "should have waited and seen how AMS settled down". He commented that,

> "One of the key issues to this project is that the MIS was designed prior to AMS going live, so nobody had a clue how AMS was going to perform or what sort of data was going to go into it, how they were going to use".

From the users point of view the timing of AMS and MIS was difficult too,

> "You have to appreciate that we were learning all sorts of other things at the same time, AMS which was huge at that time. Most managers at that time were working horrendous hours, they didn't have the time to play with the new toy or to learn the new toy, when they were getting sufficient information to make their management decisions from elsewhere".

but aren't we discussing generation?

Validity Claims

Following Mingers (1995), rightness, truth and effectiveness are the main validity claims for action. Main concerns are with a lack of change control on AMS, and the limited validation of the data model.

Rightness:

According to the participants the data in the data warehouse is relatively accurate, although they recognise that accuracy is dependent on the correct input of data at the AMS level (and hence the front line employees having a suitable understanding of the policies). In addition, there are some data integrity issues between the data in MIS and the data in AMS. For example in some circumstances a record can be physically deleted on AMS. This has flow on effects for the MIS which relies on last update flags when updating the database each week. The deleted records can only be discovered periodically by a complete refresh of the database. Participant G spoke of these change control problems,

> "I have good confidence in the application data because that's the only table that's got a last update date in AMS, and MIS gets every change made in AMS every week, except the deleted applications, there is no way MIS can get hold of them, so we have to refresh on a monthly basis. Other parts are less accurate like identities because they don't have a last change data on the fifteen million identities on AMS, so I can only pick up things that have been added, if they go back and make a minor

> change on a record that was created five years ago there is no easy way to pick that up other than refreshing the whole lot".

Truth:

is this reasonable ?

It is assumed that truth in this context was determined through a validation of the data warehouse, in particular the data model. The data model was validated pragmatically by testing the data warehouse. This is an acceptable approach used in industry. However, regrettably this may result in the implementation of a data organisation that does not necessarily meet the business requirements. A number of problems such as the lack of motivation to validate the system, were mentioned,

> "At go live time it took an awful lot of time for me to persuade them that they actually needed to test the MIS component of the system, it was small in comparison to the effort going on for the development and implementation of AMS itself".

The acceptance testing too was by no means thorough, as participant G remarked,

> "They did eventually agree to put a person on to acceptance testing, and I'm not sure it actually did them any good. Basically what they did was I gave them a cube with some stats in, I told them how a particular number was being calculated, they just went away and got the numbers of rows and re-did the same calculation and low and behold they got the same answer".

Since the initial design major changes have taken place, some of these changes may have been prevented if they had included a larger range of users on the acceptance testing (including the development of formal testing procedures). Participant G continued,

> "There was no testing as to whether that met the business need at all. ...none of the cubes that went live, that were in the design spec, and originally developed, are now in use. ...A completely different set of cubes and KPIs and ways of calculating and presenting things has been developed in the last three to four years."

Effectiveness:

In the context of this case study, effectiveness was discussed in terms of the data warehouse design. The designer said that the physical design originally involved,

> "a lot more denormalisation and consolidation of data, but we probably moved away from that for pragmatic reasons of getting the data out as quickly as we can, we use PowerPlay to calculate all the aggregations and statistics etc. So we want to get the data out of MIS and into AMS as quickly as possible".

However, four years later the designer concedes that he would design it quite differently now because "we certainly have a much better idea of what sorts of questions are asked of the data now. So I might denormalise more to improve response time".

Participant B said of the data warehouse design,

> "The design is not easy to modify, it should be simplified quite a bit and some of it is with the benefit of hindsight. ...At the moment I think the way the design works is probably not that good. It would be good to get an easy interface and improve the delivery and the content so that people want to use it and want to rely on it".

Participant G said in retrospect he would denormalise more by capturing some of the identity information on the application,

> "Any query has almost got three fact tables: application plus application/identity plus identity. ...So it all hinges around these three tables, and I've got them as three separate tables. I wouldn't necessarily do it that way now because if you want to do a join between those two tables, on this tiny box[30], it takes ages".

Action

Action may correspond in this case to the action resulting from the implementation of the model. This should imply "competence in the semantic and syntactic rules of the language or sign system" (Mingers, 1995, p.301), if the data model is to be understood. However, the problem here may be that the users do not have competence in the business rules, partly because the data is very complicated and partly because they do not understand the application process well (according to participant F). Therefore, the implication is that they may not understand the data model or the information in the data warehouse.

Despite the problems throughout this particular project, there has been some significant use of the data warehouse. For example, participant D said that initially they used MIS to validate their manual records, "when we knew ...decisions were being made from MIS and the information on MIS, it was important for me to know what was coming out on there, to understand it and to use it to help make our decisions". At the branches, they use reports provided by MIS to "give us our

[30] However, these technology issues are currently being addressed.

decisions that we have made for the week, or the month, but we tend to more focus on the month, so we can see how well we are doing in terms of targets".

Whereas participant A discussed the fact that the MIS unit currently held the responsibility for supplying the information from MIS to the business people,

> "We send them out monthly reports and we do a lot of ad-hoc queries. There used to only be one person doing that before I arrived and not everything was getting done, now there's two of us, and we've got a project starting soon to review the whole MIS. MIS is all about reporting off a data warehouse, that's where our bread and butter comes down to".

By producing queries and reports for the business people of the organisation (such as the minister, branch managers, market managers, head office people, and policy), participant A says he learns more about the data.

Participant E uses the data for research purposes, such as,

> "Identifying the characteristics of people, ...I worked on a project looking at characteristics of humanitarian category applicants, and I identified the people who I wanted to study in further depth by doing MIS report of people grouped under the humanitarian category over a twelve month period. I then used it to identify their nationality, but I needed to go further, so that gave me the numbers that I could follow up".

This participant mostly uses MIS for quantifying processes, such as making comparisons between applicant approvals and has been building up a time series over the last three years of this data. He commented that "it is getting better in terms of the depth, so I can identify trends via nationality or via permit type". He also mentioned that because he has been able to provide concrete examples of where the data has been useful the value of the system for policy and researchers has increased. However, prior to his arrival at the organisation there had not been much use of MIS for research purposes.

As mentioned earlier the small branch participants originally used MIS to validate their manual records, however, participant D said "it tends not to be used now so much though", although the reports provided by the MIS unit tell them whether they are meeting their targets for the month.

Participant C uses MIS reports at a high level for "managing the business as well as for filling out reporting requirements, so knowing what is actually happening out there in the business, using that information to be able to forecast forward, or plan how you are resourcing a particular area".

Future Requirements

The MIS unit believes that the actual MIS database does not need to change a great deal, but the interface may do. Participant A remarked,

> "It comes to the step beyond that, and that may involve some software that sits on top of MIS to make it easier, we could do some web stuff, we are talking about how we distribute information, how you get the information to the users".

In addition, he later commented that, "I would review the whole MIS and what the requirements of our various users are and then decide on the technology used to get it to them from that". Similarly, participant E suggested a "cleaner interface that doesn't get tangled up in itself, and maybe more code based rather than point and click based". Although this participant would like the design to be more 'people' based as opposed to 'application' based, "having everything about individuals easily accessible".

Participant B suggested that staff at the branches should have access to all the information that they require for operational purposes more immediately rather than waiting for a week. In addition, he said only summary information held centrally is of interest to the head office, rather than the low level detail that they are currently exposed to.

However, once again participant D said that they were satisfied with the reports generated from the MIS unit, but later remarked, "if they said we're not going to create these reports and you're going to have to sort it out yourselves we'd be saying, how do we use MIS?"

Like participants A and E, participant B suggests changing the interface to "improve the delivery and the content so that people want to use it and want to rely on it". He also believes that traffic light reporting may be more useful for the general manager. From a design perspective this participant would like to "try and engage the people who are going to use it a lot more, to say look what do your really want to get out of it, and make that drive the design, rather than from a technical point of view".

Participant C would also like to see all areas of the business connected and less reliance on manual systems, she said "make sure that it managed to encompass the whole business in a proper way". Currently, the organisation are investigating the possibility of providing MIS information via the web rather than via spreadsheets.

> "We've moved from having a client product set to having an enterprise server product set, so what they are about to do is to install PowerPlay enterprise server and start publishing data back out to the branches across the web. So everyone can now see the same source of data, there's no argument about your database is out of date, or your database is different to mine".

Most of the factors mentioned previously also apply at this level. Additional factors that contribute to the lack of 'action' are: providing information that can be interpreted, reliance on manual systems and the number of current users with access.

Mair's (1999) report describes an email survey where the respondents[31] inferred that they "did want to use information more regularly to manage their operations, but that it needed to be simple" (p.12). Also, the MIS unit talk of "getting the information to the users" as the key problem. This may well be a valid issue, but they may find that even if the information is provided in another format, such as over the web, it may not be meaningful or compelling enough to motivate the business users to use it. Participant C discussed the problems with reliance on manual systems in some parts of the organisation,

> "What it doesn't do is really help us hook up all areas of our business. Ideally we would have one system for everyone, for example our refuge side don't use this particular information, ...they tend to use manual systems which is not ideal".

Summary

The investigation of the design and use of this complex data warehouse has uncovered an array of problems relating to the generation of meaning from the physical data model and the production of the data model from meaning.

The data warehouse was developed to meet the needs of three types of user groups: senior managers, analysts in national office and branch managers (Mair, 1999). However, it is clear from the participants' comments that not all users are satisfied with the way in which the data is presented and understood. Currently, the MIS unit run queries and produce spreadsheet reports to provide information that non-

[31] The survey was to all branch managers.

technical users can understand. In fact, the majority of users depend on the MIS unit, especially the branch managers who do not directly use the data warehouse.

A number of factors that inhibit or prevent an in-depth understanding of semantic content were classified as: technical, cultural, training, resource, data or design related. One of the most prominent themes mentioned throughout this case study is the complexity of the software. Consequently, the business users experience problems with understanding the information produced by the MIS. Nevertheless, Mair (1999) says,

> "for the most part, staff weren't interested in the information, except when there was a problem. Not only was the information readily available to them, but they had at their disposal Cognos software that allowed them to simply answer most questions they might have, by using techniques to drill into the data" (p.2).

Cultural issues have been mentioned as inhibiting factors by other researchers (Mair, 1999) and the development team. However, it may be unfair to suggest that aptitude and motivation are issues at the small branch level. Because as participant D remarked, from a small branch perspective, (particularly the day-to-day operating), there is no real need for access to the raw data on MIS. In his opinion, they do not need further detail because their needs are supported by specific reports provided to the branch. Nevertheless, a training program that provides these types of users with the opportunity to use the tools in such a way that helps their day-to-day operating, may be beneficial.

The case study also described issues with the production of the data model. For example, the intention of the data modeller and the organisation was unclear from the business requirements. However, the participants noted that the motivation for the project was from a senior management perspective. In addition, the design data model was based directly on the AMS design (source database) and there was evidence of only one workable design solution.

Despite these problems, the participants described areas where the data warehouse has been useful. For example: for validating their manual records, comparing AMS with MIS data for audit purposes and identifying whether the branches are meeting their targets. In addition, the data warehouse has improved in terms of the depth of data it can now provide. Participant E uses MIS for research purposes such as identifying the characteristics of people, making comparisons between applicant approvals and has built up a time series of the data over the last three years.

8 Strategies for Achieving Semantic Integrity

> *"[W]e normally understand something by modeling it and then determining correspondences between the two domains. In some cases, we are lucky: We can, as it were, keep an eye on each domain, merging the images in our mind's eye. In other cases, notably when one of the domains is the external world, we are not so lucky..."* (Rapaport, 1998, p.74).

Introduction

The case study revealed inhibiting factors for both the generation of meaning from a data model and the production of a data model from meaning. Because of the nature of the data warehouse project the inhibiting factors are those associated with understanding the **implemented** data model. A different set of strategies may be necessary for understanding a design data model. The strategies proposed in this chapter address the factors inhibiting the generation of meaning from a data model, in particular for improving understanding. It is outside the scope of this research to describe in detail strategies for the production of a data model from meaning, this is an area for future research.

Generating Meaning from a Data Model

Strategies for Understanding

To achieve an understanding of the semantic content of the data model and to address the quality goals of 'meaningfulness' and 'comprehensibility' a number of general strategies are suggested. For example, the more ~~involved~~ the end users are involved in the design activity the more they will value and *understand* the information provided by the data warehouse. It is the responsibility of the data modeller to provide the 'tools' for achieving this. Prior to the development of the data model the designers should document contextual information by conducting interviews to determine: how the end user best acquires understanding, what their perception of the business domain is, how they intend to use the information and what reasoning they will apply when justifying the data requirements.

One of the inhibiting factors for understanding is creating a design data model that supports the different stakeholder requirements. However, Darke and Shanks (1998) propose that user viewpoint modelling provides "a means of identifying and managing the number and potential diversity of views and requirements" (p.235). They claim that explicit user viewpoint modelling is useful where a project involves,

> "diverse groups of users where the functionality of systems cross the boundaries of the various groups. Explicit user viewpoint modeling is then valuable in ensuring that all potential views and requirements are identified and represented, and it assists in managing their integration" (*ibid.* pp. 235-236).

During the design stage, the data modellers and the data consumers should have overlapping meaning structures with the motivation for reaching an agreed understanding of the model. This is dependent on the depth of knowledge the end user has of the problem domain, their ability to communicate that knowledge and their willingness to reach agreement. Kent (1978) notes the difficulty with achieving a shared view when he writes,

> "*[T]* he chances of achieving such a shared view become poorer when we try to encompass broader purposes, and to involve more people. This is precisely why the question is becoming more relevant today: the trust of technology is to foster interaction among greater numbers of people, and to integrate processes into monoliths serving wider and wider purposes. It is in this environment that discrepancies in fundamental assumptions will become increasingly exposed" (p.203).

As mentioned in chapter two a useful technique to facilitate communication and for validating the semantic content of the data model is the natural language technique NaLER (Natural Language for E-R/R) developed by Atkins and Patrick (2000), this is the recommended activity for achieving agreed understanding of the data model. The only other way to validate the model is to build the database. Atkins (2000) says of the technique,

> "it provides a way to communicate the semantic content of a design model to users and removes the necessity for users to understand the language, and the decisions, underlying the datalogical design model" (p.159).

The technique involves the creation of natural language sentences based on the components of the design data model. These sentences are populated with relevant examples taken from the existing database, which are then verified by the users. This is a useful method for communicating the semantic content of the data model because the entities, attributes and relationships alone do not have meaning, but the NaLER sentences can have meaning. From the data modeller's perspective the NaLER technique may also be a fruitful self-monitoring exercise during the data warehouse design (Atkins, 2000). Moreover, it may be useful when verifying the source database design(s) with the stakeholders who need to understand this level of detail.

Comparing the logical data model with reality can be successfully achieved using natural language. Atkins (2000) writes,

> "As NaLER provides a means of extracting a complete set of natural language sentences from an E-R/R model, any number of sets can be compared either with each other or with an initial set extracted directly from the UoD. Such comparisons can highlight incorrect or new semantics and provides a genuine basis for assessing semantic equivalence between differing data structures" (p.150).

The semantic content of the model should be verified by 'domain experts' who ensure an accurate and useful perspective of an organisation's "slice of reality" (Biller & Neuhold, 1978 p.11); therefore, it is very important that appropriate users are chosen to verify the model. The success of the verification activity depends on how well the domain experts understand the NaLER sentences. Sentences can be understood directly or indirectly, although most of our understanding may well be indirect (Lakoff, 1987, p.294). Direct understanding requires characterisations of directly understood sentences and directly understood situations. Lakoff (1987) claims that a sentence is directly understood "if the concepts associated with it are directly meaningful" (*ibid.*p.292). He portrays truth relative to direct understanding as a "correspondence between the understanding of the sentence and the understanding of the situation" (*ibid.*p.293). For example, the sentence in the zoo scenario: "Each animal is housed in one enclosure" may be directly understood as a sentence and as the situation. Lakoff (1987) fits the direct understanding of a sentence and a situation by:

- The mental image associated with the person's basic-level concept of 'animal' can accord with their perception of the overall shape of the animal.
- The mental image associated with the person's basic-level concept of 'enclosure' can accord with the person's perception of the enclosure.

- The image schemas that constitute the person's understanding of 'housed' can accord with the person's perception of the relationship between the animal and the enclosure (*ibid.* p.293).

Therefore, truth can be characterised relative to a direct understanding if "the direct understanding of the sentence is in **overall accord**[32] with the direct understanding of the situation" (*ibid.* p.293).

Nonetheless, this may not be straight forward, because the understanding of the situation must always be taken into account, and most importantly, an accurate account of 'accord with' must be determined. Therefore, a fruitful way to verify the design data model is to involve those domain experts who have directly experienced and understood the situation, and who could verify the truth content of the sentences (these domain experts may be the 'real' end users). This also has implications for the data modelling activity, in that the data modeller should try to gain an understanding of both the situation (by experiencing it) and therefore gain an understanding of the sentences that will make up the model. With regard to verifying the design data model, the sentences are declared *true* in a given situation if the person's understanding of the sentence fits their understanding of the situation closely enough for their purposes.

Furthermore, Lakoff (1987) notes that sentences and situations are often understood indirectly. A sentence is indirectly understood "if the concepts associated with it by the grammar are indirectly meaningful" (*ibid.*p.294). However, he notes that there is no detailed account of how situations are understood indirectly, but promotes a criterion of relative accuracy and good sense for situational understanding. He writes, "Such a criterion would maximize directness of understanding. It would prefer understandings that are more direct to those that are less so" (*ibid.*p.294). From this perspective a domain expert may understand a sentence as being true in a given situation if their understanding of the sentence fits their understanding of the situation closely enough for their purposes (as opposed to being an absolute truth). Lakoff (1987) concludes,

> "When there is only one conceivable understanding of a situation, then truth appears to be absolute. But when it is clear that more than one

[32] Emphasis added.

> understanding is possible, then it becomes clear that truth is relative to understanding" (p.295).

Truth is recognised by Lakoff (1987) as a human concept, "subject to the laws of human thought" (p.296). He says that there are central and noncentral truths. Central truths are described by directly understood concepts, such concepts are: basic level concepts in the physical domain or kinaesthetic image-schematic concepts (*ibid.*p.296). In this context if entities in a data model are basic level concepts (such as animal and enclosure) and relationships are a preconceptually structured experience of a kinaesthetic relation (such as inside or contain), then this makes the situation one that is directly understood. According to Lakoff (1987),

> "Central truths are true by virtue of the directness of fit between the preconceptual structure of experience and the preconceptual structure in terms of which the sentence is understood" (p.297).

Nevertheless, most sentences do not express central truths,

> "they are sentences that contain concepts that are very general or very specific or abstract or metaphorical or metonymic or display other kinds of "indirectness" relative to the direct structuring of experience" (*ibid.* p.297).

The data modeller needs to determine whether the domain expert has a direct understanding or an indirect understanding of the sentences and situation. Because NaLER sentences are of a specific format and are populated with real examples (Atkins and Patrick, 2000), a direct understanding should be reached most of the time. If a domain expert has an indirect understanding of the sentence, it does not mean that the understanding is any less true, although this type of understanding may generate further debate and refinement of the model. Lakoff (1987) remarks that most cases of truth involve indirectness, they make use of indirect understanding, for example: higher-level categories, metaphoric understanding, and abstraction. The domain expert may need to distinguish meaningful – basic level and image schematic concepts from concepts that are indirectly meaningful. He also claims that the best examples of truths are best examples of objects of knowledge. If the things people know the best "are those that can be closely related to basic-level experience" and "we get our basic knowledge of our immediate physical environments from our basic-level interactions with the environment, through perceiving, touching, and manipulating" (*ibid.* p.297), then the best domain experts are those performing daily tasks that are fundamental to the operation of the organisation.

Table 13 documents strategies that address the specific problems identified in the case study. These factors were recognised as inhibiting the understanding of the implemented (physical) data model.

Factor	Inhibiting Factors	Strategy
Technical	Complexity of the software	User consultation. Selection of easy to use languages that minimise errors and maximise user performance (Batra & Srinivasan, 1992, p.406) Ongoing training
	'Teething' Problems	Formal testing procedures, User involvement in the analysis, design and testing. Prototyping (Shanks and Darke, 1998a)
Social	Aptitude and Motivation	User viewpoint analysis (Darke and Shanks, 1997) Cultural immersion (Recognition and involvement) Visualisation – promote innovative use. (Shanks and Darke, 1998b)
Training	Inadequate training	Training protocol (Batra & Srinivasan, 1992, p.413) Provision of a range of training supporting both technical and business needs.
Resource	MIS Resource	Proactive users Quality MIS staff with a view of the wider context
Data	Comprehensiveness	Use of NaLER to describe the data model in sentences that the data consumers can understand (Atkins and Patrick, 2000). A representative from all data consumer groups participating in the data requirements definition. Forward thinking. Gather contextual information
	Format / Granularity	Prototyping. Data captured at the level of detail and format stipulated by the users.
Design	Query design	Employ a proper design practice (Date, 2000, p.705). Query plan and strategy (Batra & Srinavasan, 1992, p.400). Declare semantic constraints (Date, 2000, p.699).
	Limiting Reports	Reports designed based on the level of detail and format stipulated by the users.
	Pre-defined cubes	Determine the business need for multidimensional data. Prototyping
	Complexity of the AMS system	Formal change control procedures.
	Use of structured codes	Attributes Surrogate keys (Simsion, 1994)
	Application typing	Specialisation Generation of alternate models (Simsion, 1994)

Table 13: Strategies for Understanding[33]

A number of the participants discussed the issue of inadequate training. In particular, to carry out the day-to-day processing of applications it is essential that the business users have an in-depth understanding of legislative requirements, therefore specific training should be in place to help provide this understanding. Subsequent training should also be provided whenever amendments to legislation affect the processing of

[33] Where some of the strategies have not been further elaborated refer to the original sources for a detailed explanation.

applications. This is important because the more knowledgeable and successful the end users are at their daily tasks, the more they will be able to cope with, and understand, the data in the data warehouse. Obviously, continual support and training of the underlying source systems is also required.

Date (2000) recommends a disciplined approach to decision support design that avoids the problems with the star schema approach (refer to chapter three), and incorporates a logical design. Common design errors that Date (2000) discusses which may have contributed to some of the problems described in this case study are: duplicate rows, denormalisation (too early), star schemas, nulls (missing information), design of summary tables and multiple navigation paths.

Apparently some designers allow duplicates because the data is said to have no unique identifier, however, Date (2000) says that this arises because the "physical schema is not derived from a logical schema" and in this case "rows often have nonuniform meanings (especially if any nulls are present) – i.e., they are not all instantiations of the same predicate" (p.704). He also warns that although the practice of denormalisation may be acceptable at the physical level, it is not acceptable at the logical level. Furthermore, star schemas or dimensional models (or dimension maps) are really physical schemas, and the major problem with them is the ad hoc nature in which they are designed (*ibid.* p.705). He also claims that it is "possible (and desirable) to design in such a way as to avoid nulls" because the "resulting schemas often provide better storage efficiency and better I/O performance" (*ibid.* p.705).

Of particular relevance to the organisation in this study is the problem with the design of the summary tables, as Date (2000) highlights "users can become confused as to the meaning of summary data and how to formulate queries involving it" (p.705). For example, participant E has difficulty with constructing queries using Impromptu, "it is not so easy to find out how many people we had with work permits. ...because it is hard to construct the query". He remarked, "If you look at the actual SQL behind an Impromptu report it is incredibly confusing, it's like no SQL I've ever seen before". Date (2000) says that to avoid such problems, "all summary tables at the same level of aggregation should be designed as if they formed a database in their own right" (p.705).

He also writes of another pertinent issue to the case study, where designers (and users) incorrectly speak of there being a "multiplicity of navigational paths"[34] to some desired data. Specifically Date (2000) comments:

> "It is clear that users can become confused in such cases and be unsure as to which expression to use and whether or not there will be any difference in the result. Part of this problem can only be solved by proper user education... However, yet another part is due to designers allowing redundancies in the logical schema and/or letting the users access the physical schema directly, and that part of the problem can only be solved by proper design practice" (p.705).

Strategies for Connotation:

The social context described by Hirschheim *et al.* (1995) as the relationship between the social world and the data modelling, is represented by the meaning level - connotation. This is where the social factors such as political and cultural issues become apparent.

Participant F discussed problems the organisation has with understanding the social and cultural context within which the data is created in AMS, and then interpreted in MIS. In this particular organisation the process of accepting or rejecting an application is not purely objective, if it was, then the time spent interviewing and analysing potential clients would be less. Therefore, the socially constructed meaning of the data is that an outcome in one branch may be different to an outcome at another branch. Shanks and Corbitt (1999) studied a similar case where "an understanding of the social and political context in the" organisation[35] and "an awareness of bias" in the organisation "explains the different interpretations of results" (p.793). They say that a shared understanding of data is "an understanding or awareness of different points of view framed within the social context in which the understanding is developed" (*ibid.* p.791).

Table 14 presents strategies for the meaning level connotation.

[34] The same data can be located using different relational expressions.

[35] Their case study was of a University credit granting system.

Factor type	Inhibiting Factors	Strategy
Social	Access to data (political)	Conflict analysis (Shanks and Corbitt, 1999) Conflict resolution (Shanks and Corbitt, 1999; Shanks and Darke, 1998a)
	Capturing different user needs	User viewpoint analysis (Darke and Shanks, 1997).
	Inconsistencies (Cultural)	Training to support the application process Freedom from bias (Shanks and Darke, 1998a)
	The wider context (Cultural)	Relevance (Shanks and Darke, 1998a). Reputation of the data source (Shanks and Darke, 1998a).

Table 14: Strategies for Connotation

Particular techniques recommended for viewpoint analysis (also discussed in the previous section) by Shanks and Corbitt (1999) are: soft systems methodology (Checkland, 1981) and MEASUR (Stamper, 1992). Whereas the aim of conflict analysis techniques is to "encourage groups of stakeholders with conflicting interpretations of data to discuss their differences and develop mutual understandings of each others position" (Shanks and Corbitt, p.792).

The correctness of the data in the warehouse is dependent on the input of correct data at the source database (AMS) level. Clearly, in this case constraints should be set at AMS level to prevent possible anomalies occurring. Suggestions on how to minimise anomalies occurring are: either at the database level through the use of integrity constraints (also known as business rules) for example, preventing a user from entering a record twice on AMS, domain restrictions, or alternatively at the application level, by placing the rules in the application programs. (Refer to Date (2000[36]) for a full description).

Some participants suggested that the data in MIS be updated more regularly, because the **relevance** of the data is dependent on the ability to access and use current data. If possible the organisation need to update MIS more frequently, because as participant C points out managers who are managing their stock and applications on an hourly basis do not have the information to match.

Strategies for Intention:

Intention in the context of this case study is the individual, subjective meaning for a particular stakeholder and the implications of that meaning for action. Therefore, broadly speaking the main strategy for intention is to determine the intentions of the

[36] Or any earlier version of "An Introduction to Database Systems" by C.J. Date.

stakeholder groups. The lack of use at the branch level may not be an inhibiting factor as such because the participants interviewed at the local branch said there was no current need to use the information. Nevertheless, this may imply that the intention of the branch managers was never fully defined, and therefore the individual, subjective, meaning of the data consumers at the branch level, may not have been determined.

Producing a data model from meaning: General strategies

This is the epistemological dimension that Hirschheim *et al.* (1995) refer to as "how developers inquire into object systems and see phenomena in them" (p.21). The intention of the data modeller is based on the understanding of what has been indirectly communicated to them. Therefore, the problem is that 'meaningfulness' to the data modeller is very indirectly based on the experience of others. The intention should include determining for example what is to be modelled and why? Moreover, the intention of the data modeller should be overlapping with the intensions of the data consumers. Therefore, at the design stage the goal for intention is to reach **agreement** with the stakeholder groups by developing an overlapping meaning structure between the data modeller and the stakeholders.

As discussed in chapter two by using metonymy or reference-point reasoning (Lakoff, 1987), a data modeller may determine intention. Types of metonymy described were: social stereotypes, typical cases, ideals, paragons, generators, and salient models. Specifically, the data modeller may use metonymic models to aid understanding of the problem domain. However, development of these ideas is an area for future research.

Training should also help the end users recognise the data they need and the organisation needs so they can communicate intention. An understanding of how their data requirements contribute to the wider context may be very useful.

It is outside the scope of this research to stipulate strategies for the meaning levels generation and action. Generation is the conversion of intention into specific action, therefore generation is the data modelling activity. However, general strategies that could be refined further are:

- the development of a logical data model from which the data warehouse design is based (Date, 2000).

- stakeholders need to have sufficient understanding of the business and their job role to contribute in a useful way to the activity of data model validation.

A major strategy from a data modellers' point of view would be to undertake a logical design phase (Date, 2000), this should include the development of a logical data model. It is beyond the scope of this research to provide a detailed description of logical data modelling, this has been successfully documented by many researchers and practitioners alike (refer: Benyon, 1997; Date, 2000; Simsion, 1994; Tsrichritzis and Lochovsky, 1982). Nonetheless, the importance of such a phase is an important strategy for the meaning level generation. Another such strategy might include the generation of alternative logical models for the data warehouse.

Often data warehouse designer(s) are not in a position to redesign the operational databases, instead they have to manage the problems. This includes anticipating how short cuts or design flaws at the database level may impact on the design and success of the data warehouse. Strategies should be devised to incorporate these ideas.

In hindsight Participant G said that an organisation "shouldn't even attempt to look at the warehouse or BI side of things until the system has been in for six months", because "the users didn't even have time to think about what they need when they were trying to learn the new system".

Summary

Strategies for improving understanding have focused largely on the use of the cognitive semantics to explain how domain experts might directly or indirectly understand the design data model using the NaLER technique. Moreover, verifying the design data model should involve domain experts who have directly experienced and understood the situation, and who can verify the truth content of the sentences. Indeed, the implication for the data modeller is that where possible they should gain an understanding of both the situation through experience and thereby gaining an understanding of the sentences that will make up the model.

Other strategies proposed include the use of **proper** design techniques for data warehouse design, including the development of a logical data model (Date, 2000). As mentioned earlier, strategies for the production of a data model from meaning have not been described in detail however, some initial ideas have been formulated.

9 Conclusions and Future Research

"The practice of decision support is, regrettably, not as scientific as it might be; often, in fact, it is quite ad hoc. In particular, it tends to be driven much more by physical considerations than by logical ones (indeed, the distinction between physical and logical matters is often very blurred in the decision support environment)" (Date, 2000, p.695).

Reflection

Meaning is *generated* from information by interpreters through a process of digitalisation that extracts only some of the information available (Mingers, 1995). Based on the findings from this research intersubjective *meaning* defines the semantic integrity of the data warehouse. Similarly Dretske (1998) remarks,

> "The distinction between meaning and information, between belief and knowledge, is a distinction that only makes sense from the outside. But what makes sense *only from the inside* of the machine or person whose behavior is being explained cannot (according to this way of thinking) help to explain that machine's (or person's) behavior" (p.270).

The main purpose of this research was to explore the importance of semantic integrity during data warehouse design and its impact on the successful use of the implemented warehouse. This was investigated through the use of a single detailed case study.

The research propositions were:

3. Semantic integrity is an important critical success factor in determining the effectiveness of a data warehousing project.
4. A 'good' data model is an important critical success factor in determining semantic integrity.

The first proposition would appear realistic because as discussed in chapter two a user can only reach points of action, knowledge or wisdom if they have sufficient understanding of the semantic content of the data model. This was evident from the case investigated where the participants discussed problems with interpreting the data complexity. For example, participant C commented that she was confident in the quality of the data in the data warehouse, but she noted there are "issues around interpreting the data complexity". She continued,

> "It comes back to the complexity of the data you have to have someone interpreting it well and knowing what it actually means".

Mair (1999) also said that although the information was readily available to the staff, they were not interested in the information, except when there was a problem. Accordingly, understanding the semantic content of the data model is a prerequisite for action and also an important factor in determining the effectiveness of a data warehouse.

From the participants comments the data cubes do not assist with providing a detailed understanding of the semantic content of the information in the data warehouse. The~~y~~ cubes are only useful for providing summary information, but as participant B remarked sometimes even that summary data can be meaningless.

The fundamental reason for examining meaning in this research has been to address the need for quality in data modelling. The second proposition recognises this and appears acceptable at face value. However, a 'good' data model may be created through 'good' design practice, although, the information represented by the model may never be wholly, semantically accurate. As Kent (1978) succinctly writes,

> "Data structures are artificial formalisms. They differ from information in the same sense that grammars don't describe the language we really use, and formal logical systems don't describe the way we think" (p.v).

This research suggests that a data modelling activity, can never, fully guarantee the semantic integrity of a data warehouse, although the process of undergoing a data modelling exercise may be an essential factor when determining semantic accuracy. Hay (1997) also argues that a data modelling exercise is necessary when he comments,

> "Contrary to what many believe, the multi-dimensional model does not replace the conceptual, relational model. Indeed, the development of a normalized conceptual model will be **critical**[37] to the **success**[1] of any data warehouse project. It is impossible to manage a set of data marts without one" (p.7).

[37] Emphasis added.

Furthermore, the organisation clearly did not predict the implications of validating the design data model and the impact this might have had on the success of a data warehouse project. Participant G remarked that the organisation had to be persuaded to test the implementation model and while they agreed, the testing was very superficial and insufficient.

The analysis of this case study has shown a diverse range of stakeholder groups within the organisation. However, the data warehouse design and the tool set used to access the information do not support their needs. Also, from a small branch perspective there is no motivation to understand the data warehouse any more than they currently do. This may imply that the intention of the branch managers was never fully defined during the design stage. However, if their requirements change in the future, practising the strategies described in chapter eight may aid their understanding of the data warehouse semantic content.

Participant A commented that, from an organisational perspective, the problem is "more the gap between getting the information out to the people that need it", he remarked that the business people need information to do their jobs and make decisions, but there is a gap between having the data and getting it to the end users. He continued, "so that's our job to get the information from the data warehouse out to the people that need to use it". However, they may endeavour to provide this information to more end users by using a different strategy, but according to participant B the "content is not compelling enough". Therefore, even if the information is provided in a different format, perhaps using a different tool set, this may not address the larger problem of ensuring the data is compelling and meaningful.

A consultant who also worked for the company for ten years discussed the complexity of data and knowledge required of the people, for example to successfully process an application, there is a "huge area of knowledge that one has to acquire" and "it's not something that is picked up, it takes a long time". Consequently, before further investigation is made into "getting the information out" it may be fruitful for the organisation to determine why the end users, in particular, the branch managers do not use the data warehouse. For example, participant A alludes to reasons why the data warehouse may not be used at the branch level when he comments,

> "The reality was that many people did not have the aptitude, or could not work out the product, or could not be bothered, whatever, and that nobody uses it – nobody in the branch uses it. So, for whatever reason, that leaves a lot of people in the situation where they don't have access to the information, either because they don't use it, or because the product is too hard or whatever, some people would say that's their fault. I believe that it's the responsibility of MIS to manage that – it should be easy".

They need to determine the **actual** reasons for the lack of use at the branch level, either by asking the branch managers through interviewing or sending out a questionnaire requesting specific information about their perceived data needs. Participant A recognises that there was a problem with addressing the users requirements, however, while he regards information distribution as the key issue, it may be also be useful to explore the larger problem of understanding the data complexity. Currently the organisation is looking for a new way to provide the information. However, the tools are only a *means* to the end, changing the tools may not solve the problem of understanding the semantic content of the information.

Shortcomings

A potential criticism of this research might be that no sophisticated data analysis techniques were utilised. This was partly due to time constraints and also due to the type of data that was collected. If time had permitted more participants should have been interviewed, for example a larger sample at the branch level (unfortunately it was difficult to gain access to just one branch and their employees). However, such involvement may have allowed a comparison of views across branches, which may have substantiated the findings.

It may also have been useful to compare the factors discovered as a result of this case with other data warehousing projects. However, this particular organisation and the problems they are experiencing with their data warehouse were particularly illuminating from the perspective of this research purpose. The framework proposed for understanding meaning in data modelling worked well for describing this case, but further empirical studies need to be undertaken to test the efficacy of the framework in other settings.

Future Work

Throughout this report several references have been made to areas for future research. For example, an interesting question to investigate might be: how is meaning incorporated in the activity of data modelling? This was not documented as part of this case study because most of the case study participants were not involved in the data modelling activity. Also, the generation of meaning from a data model was determined by discussing the meaningfulness of the physical data model. The framework proposed in chapter two might be particularly useful for examining intersubjective meaning at the logical or design data modelling stage.

Chapter two also provides several avenues for future research especially incorporating the work of Lakoff (1987) for example, the theory of categorisation and the notion of prototype effects may be a fruitful area for achieving semantic integrity. Also investigating the use of a natural language approach for verifying the semantic integrity may be rewarding. The research process also provided key areas for future research such as developing practical guidelines and critical appraisal guidelines for multiple case study research. Moreover, the pilot case study preparation also resulted in four classifications of data warehousing projects (refer chapter six), these types of projects may be useful to study and compare using the framework presented in chapter two.

Overall this exploration has revealed the *vital* importance of semantic integrity during data warehouse design and also the *significant* impact this has on the successful use of the implemented warehouse.

REFERENCES

ANSI (AMERICAN NATIONAL STANDARDS INSTITUTE) (1975): ANSI/X3/SPARC Study Group on Data Base Management Systems: Interim Report, ACM SIGMOD Bulletin:7(2).

ARTZ, J. M. (1997): How Good is that Data in the Warehouse? The Database for Advances in Information Systems, 28(3): 21-31.

ATKINS, C.F. (2000): INTECoM: *An integrated conceptual data modelling framework*, Unpublished Thesis, Department of Information Systems, Massey University, New Zealand.

ATKINS, C. and PATRICK, J. (2000): NaLER: A natural language method for interpreting entity-relationship models, *Campus-Wide Information Systems,* 17(3): 81-84.

BATRA, D. and SRINIVASAN, A. (1992): A review and analysis of the usability of data management environments, *International Journal of Man-Machine Studies*, 36: 395-417.

BENBASAT, I., GOLDSTEIN, D.K., and MEAD, M. (1987): The case research strategy in studies of information systems, *MIS Quarterly*, 11(3): 368-386.

BENYON, D. (1997): *Information and Data Modelling* (2nd Edition), McGraw-Hill, London.

BENYON-DAVIES, P. (1992): The realities of database design: an essay on the sociology, semiology and pedagogy of database work, *Journal of Information Systems*, 2: 207-220.

BILLER, H., and NEUHOLD, E. (1978): Semantics of Data Bases: The Semantics of Data Models, *Information Systems*, 3: 11-30 (quoted in ATKINS, C. F., 2000).

BOGDAN, R. and BIKLEN, S.K. (1992): *Qualitative research for education: An introduction to theory and methods* (2nd edition), Boston, Allyn & Bacon, (quoted in MILES, M.B and HUBERMAN, A.M. 1994).

BOLAND, R.J. (1987): The In-formation of Information Systems, in: BOLAND, R.J. and HIRSCHHEIM, R. (eds.). *Critical Issues in Information Systems.* Wiley, Chichester.

BOLAND, R.J. (1991): Information Systems Use as a Hermeneutic Process, In NISSEN, H-E. KLEIN, H.K. and HIRSCHHEIM R.A. (eds.), *Information Systems Research: Contemporary Approaches and Emergent Traditions*, North-Holland, Amsterdam: 439-464.

BRITTEN, N., JONES, R., MURPHY, E. and STACY, R. (1995): Qualitative research methods in general practice and primary care, *Family Practice*, 12: 104-14, (quoted in GREENHALGH, T. and TAYLOR, R. 1997).

BRONTS, G., BROUWER, S.J., MARTENS, C.L.J. and PROPER, H.A. (1995): A Unifying Object Role Modelling Theory, *Information Systems*, 20(3): 213-235.

BRYSON, B. (1989): *The Lost Continent – Travels in Small Town America*, ABACUS, Great Britain.

BURMEISTER, O.K. (1995): Evaluating the Factors that Facilitate Deep Understanding of Data Analysis, *Australian Journal of Information Systems*, 3(1): 2-13.

CANNON, J. (1998): Making Sense of The Interview Material: Thematising, NUD*IST, and 10meg of Transcripts, *Proc. Annual Meeting of the Australian Association for Research in Education, Adelaide, Nov-Dec.*

CARROLL, J.M. and SWATMAN, P.A. (2000): Structured-case: A methodological framework for building theory in information systems research, *Proc. European Conference on Information Systems,* Vienna, July 3-5th, 116-123.

CATTELL, R.G.G. (1994): *Object Data Management: Object-Oriented and Extended Relational Database Systems (revised edition),* Addison Wesley, Reading, MA (quoted in ATKINS, C.F., 2000).

CAVAYE, A.L.M. (1996): Case study research: a multi-faceted research approach for IS, *Information Systems Journal,* 6: 227-242.

CHECKLAND, P.B. (1981): *Systems Theory, Systems Practice*, Wiley, Chichester.

CHEN, P.P. (1976): The Entity-Relationship Model - Toward a Unified View of Data, *ACM Transactions on Database Systems*, 1(1): 9-36.

CODD, E.F. (1970): *A Relational Model of Data for Large Shared Data Banks.* Communications of the ACM 13 (6): 377–387.

CREASY, P. and MOULIN, B. (1992): Adding Semantics to Semantic Data Models, (quoted in Nagle et al., 1992).

DARKE, P. and SHANKS, G. (1997): User viewpoint modelling: understanding and representing user viewpoints during requirements definition, *Information Systems Journal*, 7: 213-239.

DARKE, P., SHANKS, G. and BROADBENT, M. (1998): Successfully Completing Case Study Research: Combining Rigour, Relevance and Pragmatism, *Information Systems Journal*, 8: 273-289.

DATE, C.J. (2000): *An Introduction to Database Systems (7th Edition),* Addison-Wesley, Reading, Massachusetts.

de CARTERET, C. and VIDGEN, R. (1995): *Data Modelling for Information Systems*, Pitman Publishing, London.

DEVLIN, B. (1997): *Data Warehouse from Architecture to Implementation*, Addison Wesley, Reading, Massachusetts.

DENZIN, N.K. and LINCOLN, Y.S. (eds.)(1994): *Handbook of qualitative research,* London, Sage (quoted in GREENHALGH, T. and TAYLOR, R. 1997).

DRETSKE, F. (1981): *Knowledge and the Flow of Information*, Basil Blackwell, Oxford.

DRETSKE, F. (1998): Putting Information to Work, In: *Artificial Intelligence and Cognitive Science Conceptual Issues,* CLARK, A. and Toribio, J. (eds.), Garland, New York.

EISENHARDT, K. (1989): Building Theories from case study research, *Academy of Management Review*, 14(4): 532-550 (quoted in WALSHAM, G. 1995).

ENGLISH, L. P. (1999): *Improving Data Warehouse and Business Information Quality*, Wiley, New York, USA.

FINKELSTEIN, C. (1989): *An Introduction to Information Engineering: From Strategy Planning to Information Systems*, Addison Wesley, Sydney.

FLYNN, D. (1998): *Information Systems Requirements: Determination and Analysis (2nd edition),* McGraw-Hill, London.

FONG, J. ZENG, X. (1997): Data Warehouse for Decision Support, In: *Data Mining, Data Warehousing & Client/Server Databases*, Proc. 8th International Database Workshop, Hong Kong, July, 195-207.

FORESTER, J. (1992): Critical Ethnography: On Field Work in an Habermasian Way, In *Critical Management Studies*, ALVESSON, M. and WILLMOTT, H. (eds.), Sage Publications, London, (quoted in KLEIN, H. K. and MYERS, M. 1999).

GALLIERS, R.D. (1993): Research Issues in Information Systems, *Journal of Information Technology*, 8(2): 92-98.

GIANNOCCARO, A., SHANKS, G. and DARKE, P. (1999): Stakeholder Perceptions of Data Quality in a Data Warehouse Environment, *Australian Computer Journal*, 31(4): 110-117.

GLASER, B. G., and STRAUSS, A.L. (1967). *The discovery of grounded theory: strategies for qualitative research*, Chicago, Aldine (quoted in MILES, M.B. and HUBERMAN, A.M 1994).

GREENHALGH, T. (1997): Papers that go beyond numbers (qualitative research), In GREENHALGH, T. (ed.), *How to read a paper: The basics of evidence based medicine*, BMJ Publishing Group, UK, 151-162.

HABERMAS, J. (1984): *The Theory of Communicative Action, Vol 1: Reason and the Rationalization of Society*, Heinemann, London

HAMMER, M. and MCLEOD, D. (1981): Database Description with SDM: A Semantic Database Model, *ACM Transactions on Database Systems*, 6(3): .351-386.

HAY, D. (1997): From a Relational to a Multi-Dimensional Data Base, Oracle Developer Tools User Group Conference, www.essentialstrategies.com.

HIRSCHHEIM, R., KLEIN, H.K. and LYYTINEN, K. (1995): *Information Systems Development and Data Modelling, Conceptual and Philosophical Foundations*, Cambridge University Press, Cambridge.

HITCHMAN, S. (1995): Practitioner perceptions on the use of some semantic concepts in the entity-relationship model, *European Journal of Information Systems*, 4: 31-40.

INMON, W. H. (1993): *Building the Data Warehouse,* Wiley-QED, New York.

INMON, W.H., IMHOFF, C. and SOUSA, R. (1998): *Corporate Information Factory,* Wiley, New York.

KAHN, B., STRONG, D.M. and WANG, R.Y. (1997): A Model for Delivering Quality Information as Product and Service, *Proceedings of the 1997 Conference on Information Quality*, Massachusetts Institute of Technology: 80-94, (quoted in: SHANKS, G. and DARKE, P. 1998a).

KENT, W. (1978): *Data and Reality,* North-Holland, Amsterdam.

KIMBALL, R. (1996): *The Data Warehouse Toolkit*, Wiley, New York.

KIMBALL, R., REEVES, L., ROSS, M., THORNTHWAITE, W. (1998): *The Data Warehouse Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, Wiley, New York.

KLEIN, H. and MYERS, M. (1999): A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems, *MIS Quarterly*, and 23(1): 67-93.

KOSKO, B. (1993): *Fuzzy Thinking: the new science of fuzzy logic*, Harper Collins, London.

KROGSTIE, J., LINDLAND O.I and SINDRE, G. (1995): Towards a deeper understanding of Quality in Requirements Engineering, *Proceedings of 7th CAiSE, Jyvaskyla, Finland.* June: unnumbered, (quoted in: ATKINS, C.F. 2000).

LAKOFF, G. (1987): *Women, Fire and Dangerous Things*, University of Chicago Press.

LAWRENCE, S. (1997a): Unpublished Project Report. (This report may be accessed through the Information Systems Department at Massey University, Palmerston North upon signing a confidentiality agreement).

LAWRENCE, S. (1997b): Unpublished Project Report. (This report may be accessed through the Information Systems Department at Massey University, Palmerston North upon signing a confidentiality agreement).

LEWIS, P. (1991) The decision making basis for information systems: the contribution of Vickers' concept of appreciation to a soft systems perspective, *European. Journal of Information Systems*, 1: 33-43. (quoted in: MINGERS, J.C. 1995).

LINDLAND, O.I., SINDRE, G and SØLVBERG, A. (1994): Understanding Quality in Conceptual Modelling, *IEEE Software*, March: 42-49, (quoted in SHANKS, G. and DARKE, P. 1998a).

LOFLAND, J. (1971): *Analyzing social settings: A guide to qualitative observation and analysisI,* Belmont, CA: Wadsworth (quoted in MILES, M.B and HUBERMAN, A.M. 1994).

LYYTINEN, K.J. and KLEIN, H.K. (1985): The Critical Theory Of Jurgen Habermas As A Basis For A Theory Of Information Systems, in MUMFORD, E et al. (eds.), *Research Methods in Information Systems*, Elsevier Science Publishers B.V. (North-Holland): 219-235.

MAIR, P. (1999): Unpublished Report, Victoria University, Wellington, New Zealand (This report may be accessed through the Information Systems Department at Massey University, Palmerston North upon signing a confidentiality agreement).

MAKINS, M. (ed.)(1995): *Collins Concise Dictionary* (third edition), HarperCollins, Great Britain.

MARCHE, S. (1993): Measuring the stability of data models, *European Journal of Information Systems*, 2(1) 37-47.

MATTISON, R. (1996): *Data Warehousing Strategies, Technologies and Techniques,* McGraw-Hill, New York.

MATURANA, H. (1975): Representation and Communication Functions, *BCL #57/5, Biological Computer Library,* Illinois University, Urbana (quoted in MINGERS, J.C. 1995).

MATURANA, H. (1978): Biology of Language: the epistemology of language, In: MILLAR, G. and LENNENBERG, E. (eds.), *Pyschology and Biology of language and Thought: Essays in Honour of Eric Lenneberg*, (quoted in MINGERS, J.C. 1995).

MAXWELL, J.A (1996): *Qualitative research design: an interactive approach,* Thousand Oaks, CA:.Sage.

MAYS, N. and POPE, C. (1996): *Qualitative Research in Health Care.* London, BMJ Publishing Group (quoted in GREENHALGH, T. and TAYLOR, R. 1997).

McELREATH, J. (1998): Data Warehouse Design: Issues in Dimensional Data Modeling, In THURAISINGHAM, B.(ed.), *Handbook of Data Management*, CRC Press LLC, Auerbach, Boca Raton, Boston. 527-540.

McKAY, J. MARSHALL, P. (2000): Quality and Rigour of Action Research in Information Systems, *Proc. European Conference on Information Systems,* Vienna, July 3-5th, 108-115.

MILES, M.B and HUBERMAN, A.M. (1994): *Qualitative Data Analysis: An Expanded Sourcebook*, CA:.Sage.

MINGERS, J.C. (1995): Information and meaning: foundations for an intersubjective account, *Information Systems Journal*, 5: 285-306.

MINGERS, J.C. (1995b): An evaluation of Theories of Information with Regard to the Semanic and Pragmatic Aspects of Information Systems, *Warwick Business School Research Papers,* Warwick Business School, 136.

MOODY, D. and BUIST, A. (1999): Improving the Links Between Information Systems Research and Practice – Lessons from the Medical Profession, *Proc. 10th Australian Conference on Information Systems, Wellington, New Zealand,* 645-659.

MOODY, D. and SHANKS, G. (1994): What Makes a Good Data Model? Evaluating the Quality of Entity Relationship Models, *Dept of Information Systems, Monash University Working Paper Series* 12/94, Melbourne, Australia (quoted in: ATKINS, C.F. 2000).

MOODY, D. and SHANKS, G.G. (1998): What Makes a Good Data Model? A Framework for Evaluating and Improving the Quality of Entity Relationship Models, *Australian Computer Journal,* August, 30(3): 97-110.

MYERS, M. (1994): A Disaster for Everyone to See: An Interpretative Analysis of a Failed IS Project, *Accounting, Management and Information Technologies,* 4(4): 185-201, (quoted in KLEIN, H.K. and MYERS, M. 1999).

NAVATHE, S.B. (1992): Evolution of Data Modelling for Databases, *Communications of the ACM,* 35 (9): 112-123.

NEWCUM, J. (2000): *Personal Communication,* Bank One, Columbus, Ohio.

ORLIKOWSKI, W.J. (1991): Integrated Information Environment or Matrix of Control? The contradictory Implications of Information Technology, *Accounting, Management and Information Technologies,* 1(1): 9-42, (quoted in KLEIN, H.K. and MYERS, M. 1999).

PATTON, M.Q. (1990): *Qualitative evaluation and research methods* (2nd edition),Newbury park, CA:.Sage.

PUTMAN, H. (1975): The Meaning of "Meaning", in: GUNDERSON, K. (ed.), *Language, Mind and Knowledge*, Burns & Maceachern Ltd, Canada.

RAPAPORT, W. J. (1998): Syntactic Semantics and Computational Cognition. In: (eds) CLARK, A. and TORIBIO, J: *Artificial Intelligence and Cognitive Science Conceptual Issues,* Garland Publishing, New York.

RICHARDS, L. (1997): Computers and Qualitative Analysis, In KEEVES, J.P. *Educational Research, Methodology, and Measurement: An International Handbook* (2nd edition), 286-290.

SHANKS, G., O'DONNELL, P. and ARNOTT, D. (1997) Data Warehousing: A Preliminary Field Study, *Proc. 8th Australasian Conference on Information Systems*, University of South Australia, Adelaide, pp 350-365

SHANKS, G. and DARKE, P. (1998a): Understanding Data Quality in a Data Warehouse, *The Australian Computer Journal*, 30(4): 122-128.

SHANKS, G. and DARKE, P. (1998b) Incorporating Context to Improve Understanding of a Data Warehouse, *Proc. IFIP Working Group 8.3 Conference on Decision Support Systems*, Bled, Slovenia (July)

SHANKS, G and CORBITT, B. (1999): Understanding Data Quality: Social and Cultural Aspects, *Proc. 10th Australian Conference on Information Systems, Wellington, December*, 789-797.

SILVERSTON, L., INMON, I. and GRAZIANO, K. (1997): *The Data Model Resource Book: A Library of Logical Data Models and Data Warehouse Designs,* Wiley, New York.

SIMSION, G., (1994): *Data Modelling Essentials Analysis, Design and Innovation*, Van Nostrand Reinhold, Boston.

STAMPER, R. (1987): Semantics, In: BOLAND, R.J. and HIRSCHHEIM, R. (eds.), *Critical Issues in Information Systems*, Wiley, Chichester.

STAMPER, R. (1992): Signs, Organisations, Norms and Information Systems, *Proc. 3rd Australian Conference on Information Systems*, Wollongong.

STAKE, R.E. (1994): Case Studies, In DENZIN, N.K. and LINCOLN, Y.S. (eds.) *Handbook of qualitative research,* London, Sage, 236-237.

STERLING, S. (1998): Essential Modeling Options, Teradatareview, www.teradatareview.com, Fall: 30-39.

TAYI, G. and BALLOU, D. (1998): Examining Data Quality, *Communications of the ACM*: 41(2): 55-57, (quoted in: SHANKS, G. and DARKE, P. 1998a).

TOZER, G.V. (1999): *Metadata management for Information Control and Business Success*, Artech House, Boston.

TRAVIS, J. (1999): Exploring the Constructs of Evaluative Criteria for Interpretivist Research, *Proc. 10th Australian Conference on Information Systems, Wellington, December*, 1037-1048.

TSICHRITZIS, D.C. and LOCHOVSKY, F.H. (1982): *Data Models,* Prentice Hall, Englewood Cliffs, New Jersey.

TUOMI, I. (1999): Data Is More Than Knowledge: Implicatiosn of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory, *Journal of Management Information Systems,* Winter, 16(3): 103-117.

TYPANSKI, R. E. (1998): Creating a Data Warehouse within an Information Environment Architecture, in THURAISINGHAM, B.(ed.), *Handbook of Data Management,* CRC Press LLC, Auerbach, Boca Raton, Boston. 563:578.

WALSHAM, G. (1993): *Interpreting Information Systems in Organizations*, Wiley, Chichester, England.

WALSHAM, G. and WAEMA, T. (1994): Information systems strategy and implementation: a case study of a building society, *ACM Transactions on Information Systems*, 12(2): 150-173.

WALSHAM, G (1995): Interpretive case studies in IS research: nature and method, European Journal of Information Systems, 4(2): 74-81.

WHEELER, D. (2000): "The Precise Endeavour" Unpublished Research Report, Department of Information Systems, Massey University, New Zealand.

YIN, R.K. (1984): *Case Study Research: Design and Methods*. Beverly Hills, CA.: Sage.

Glossary

comprehensibility: The degree to which the structure or semantic content of the data warehouse is understandable. Understanding the semantic content is dependant on the stakeholders stock of tacit knowledge.

conceptual data model: A model, or collection of models, that records the information requirements of a system with no consideration of the specific technology by which it will be implemented. Sometimes referred to as the 'logical model' of the system (Atkins, 2000).

data mart: A data mart, which is sometimes referred to as a data cube, is a grouping of data in a read-only database according to either the specific business unit, for example a sales data mart, or the geographical location. Data marts may or may not be combined with other data marts to form a complete data warehouse. The main difference between a data mart and an enterprise data warehouse is scope.

data warehouse: A data warehouse can be either: a central database organised according to a corporate data model, or one or more other databases (data marts) organised to meet specific retrieval requirements. Decision support data is usually collected from a variety of disparate systems and kept in a data store (the 'warehouse') of its own on a separate platform.

data model: A representation of a particular set of data according to a set of conventions used to represent a simplified, formal and highly abstracted view of data (Atkins, 2000).

data modelling: A knowledge creating process that results in the creation of data model using a particular data modelling approach.

data modelling approach: A data model can be classified according to the approach adopted to create the model and how that is communicated (usually used to describe a generic type), for example the E-R approach, relational approach, OO approach, or the star-schema approach.

design data model: A formal description of all identified data requirements for a system under development (Atkins, 2000). For example, a normalised relational model.

integrity: The quality of being unimpaired; soundness, or the degree of unity; wholeness (Makins, 1995).

meaningfulness: The degree to which the content of the data warehouse is useful, fruitful or significant to a particular community of users. Meaningfulness is derived from using information in such a way that is useful for knowledge creation and decision-making.

meta data: Data that describes and characterises other data

physical data model: A representation of a data model that has been adapted for implementation purposes on a specific DBMS. This includes all the commands required to create all the database objects as well as the specification for those internal objects peculiar to the physical database (Atkins, 2000).

physical schema: The blueprint of the database as it implemented in the DBMS. The physical schema is the application of the physical data model.

semantics: The issue of 'meaning' that is, the relationship between signs and what they are supposed to represent (Stamper, 1987).

semantic integrity/accuracy: The degree to which semantic content of the information carried by the data warehouse overlaps with the meaning structures of data consumers (end users). Furthermore, data can be defined as semantically accurate from two views, by assessing the actual data values and by assessing the intended data values. For example, the actual data values can be validated using domain restrictions, however, the intended data values may only be validated through business people deriving knowledge or wisdom from that data. This research is mostly focused on the 'meaningfulness of the intended data values in the context of data warehousing.

APPENDICES

Appendix 1
Pilot Case Study Questions

The following is the list of initial questions were of the chief financial officer during the pilot case study.

Preparatory Interview Questions:

1. Do you have an enterprise wide data model (or any data model)? (Project Manager) This will look at determining what basis the organisation had to work from. (To determine whether the existence of an enterprise wide data model would make the modelling of the data warehouse easier).

2. Do you have a physical data model? For example, does the model reflect how the tables in the database are currently structured? (Project Manager) To determine how important data modelling is perceived to be by the organisation (whether this is an important step when commencing a data warehouse project).

3. How often is the data model referenced? Do the users understand the model? (Project Manager, Users). To determine how useful the data model is perceived to be. This may also reflect how current the model is. To establish user involvement and understanding.

4. How long ago was the data model constructed? (Project Manager)

5. Who was involved in developing the data model? (Project Manager, Users). To determine whether an outside source was involved and to determine their modelling knowledge/experience.

6. Is this data model maintained by anyone? (Project Manager). To determine whether a formal job role exists to support the database(s), and to determine how current the model is. Also to determine how formal the data model maintenance process is.

7. What methodology was employed by your organisation to undertake the construction of the data warehouse? (Project Manager). To identify whether a formal method was used for the development.

8. What made you decide to choose this method? (Project Manager)

9. Did you undertake any research as to what methodologies have been used successfully by other companies? (Project Manager). Because there is no definitive data warehouse methodology it may be useful to determine what (if any) research the organisation undertook.

10. At what stage in the development process did you start data modelling? (Project Manager) To determine the importance of data modelling and the relationship to the data warehousing literature.

10. What data modelling approach did you use? (For example, star schema, relational data modelling or both). (Project Manager). Can you provide description of the steps taken to construct the data model for the data warehouse?

12. What exposure have you had to the literature on data warehouse design? (e.g. Inmon, Kimball etc). Had you read or have you since read anything on data warehousing in any industry - computer journals, for example?(Project Manager)

13. How was the enterprise wide data model used as input to the data warehouse design? (Project Manager)

14. Who was responsible for the development of the warehouse data model? (Project Manager).

15. To what degree were the users involved in the data modelling? (Project Manager, Users)

16. How did you incorporate time in your data model? (ie. Temporal data modelling). (Project Manager). The literature has shown that temporal data modelling techniques are required in at least two important phases of the dw modelling process. These are: when dealing with temporal aspects of dimensions (slow varying) and secondly when the historical model for the corporate data warehouse is developed.

17. How successful has the design proven to be? i.e. how easy might it be to modify the data warehouse? (Project Manager, Users). (This may be difficult for the interviewee to answer depending on the stage at which the project is at. Often the success of a design can not be measured for some years).

18. Is the data easy to access and understand? (Project Manager, Users). To measure the effectiveness of data warehouse structure. The answer to this will depend on whether the interviewee is actively involved in the use and construction of queries.

19. How fast is the query response time? (Project Manager, Users). To measure the effectiveness of data warehouse structure. The answer to this will depend on whether the interviewee is actively involved in the use and construction of queries.

20. Do you have confidence in the quality of the data in the warehouse? (Project Manager, Users). To provide a measure of the success of the data warehouse implementation.

Appendix 2
Technical Pilot Case Questions

1. What was your role on the project?
2. At what stage of the development did you become involved?
3. Did they have a data dictionary?
4. What (if any) existing design did they use as a basis for the warehouse design.
5. Did they have any form of conceptual data model? (or any data model)?
6. Did they have a physical data model? For example, does the model reflect how the tables in the database are currently structured?
7. What methodology was used to develop the data warehouse?
8. Can you provide a brief description of the main steps involved in this methodology?
9. At what stage in the development process did you start data modelling?
10. What data modelling approach did you use? (For example, star schema, relational data modelling or both).
11. Can you provide a description of the steps taken to construct the data model?
12. How did you validate the data model once finished? Is this typical of all such projects?
13. Were the users involved in the validation process?
14. Who was involved in developing the data model? Who was responsible for the development of the warehouse data model?
15. To what degree were the users involved in the data modelling?
16. How involved do you believe the users need to be during the data modelling stage?
17. How often is the data model referenced? As far as you are aware, do the users understand the model?

18. Did you use some sort of tool to create the model?

19. Is this data model maintained by anyone?

20. How did you incorporate time in your data model? (time dimension?)

21. Did you carry that over into the other model(relational)?

22. How successful has the design proven to be? i.e. how easy might it be to modify?

23. Is the data easy to access and understand?

24. How fast is the query response time?

25. Do you have confidence in the quality of the data in the warehouse?

26. What would you do differently if you could re-design the data warehouse?

27. How do you develop a warehouse model/dimension map if there is no existing data model to work from?

28. Have you worked on projects that involve working from a database with inherent design problems and contain "dirty data"? How do you then build/design a data warehouse without also including (by default) those same design flaws?

29. What exposure have you had to the literature on data warehouse design? (e.g. Inmon, Kimball). Had you read or have you since read anything on data warehousing in any industry, computer journals, for example?

Appendix 3
User Case Study Questions

User Background

1. What is your job title and what does your job involve?
2. How long have you worked for NZIS?
3. What qualifications do you have at the tertiary level? (data modelling?)

Existing database(s) and Business Requirements (AMS):

1. Do you use AMS as part of your job? How reliable is AMS?
2. Were you involved in the original business requirements definition for AMS? (specifically the data requirements). For example, were you involved in the data modelling activity for AMS? How were you involved?
3. Do you understand and reference the technical documentation, in particular the data model?
4. Were you involved in validating the data in AMS? How was this achieved? How difficult did you find this validation process?
5. Do you feel that the validation process was adequate?
6. Were you involved in the general testing of AMS? How difficult did you find the testing?

Data Warehouse (MIS - use and validation):

1. Do you use MIS as part of your job? How reliable is MIS?
2. What kind of questions do you ask of the data cubes?
3. Were you involved in the original business requirements definition for MIS? (specifically the data requirements). For example, were you involved in the data modelling activity for MIS? How were you involved?
4. Do you have a copy of the technical documentation for MIS?
5. Do you understand the technical documentation, in particular the data model?
6. Do you reference the technical documentation, in particular the data model?

7. Were you involved in validating the data in MIS? How was this achieved? How difficult did you find this validation process?

8. Do you feel that the validation process was adequate?

9. Were you involved in the general testing of MIS? How difficult did you find the testing?

MIS Project Outcome:

1. How has the experience of using MIS been for you on a day to day basis?

2. How has the experience of using MIS been for NZIS?

3. Does the data support your business requirements? What changes would be needed to fully support your needs?

4. Is the data represented in a format that is useful to you. That is, is it at the correct level of detail (granularity)?

5. Is the data in the data cubes meaningful and useful to you? Can you provide an example where the data is meaningful, and an example where it is not?

6. Do you have confidence in the quality of the data in the warehouse?

7. Do you feel you have a good understanding of the data in the data cubes?

8. Is the data correct?

9. Is the data unambiguous?

10. Is the data comprehensive? (the data cubes contain all the necessary data).

11. Do you get meaningful answers to queries? (never, very occasionally, sometimes, always) Why?

12. Is the data easy to access and understand? (Can you retrieve the data in a manner that makes sense to you?)

13. How fast is the query response time?

14. Are you able to produce all the reports/analyses you require?

15. What type of action results in your using MIS, or knowledge do you gain from using MIS?

16. What types of changes would you ask to be made of MIS if it was to be re-designed?

17. Has using MIS altered in any way how you do your job?

18. Has using MIS altered in any way how the organisation works?

Appendix 4
Technical Case Study Questions

General:

1. Could you please provide a brief description of the project (main inspiration).
2. What was your role on the project?
3. At what stage of the development did you become involved?

Background

1. What qualifications do you have at the tertiary level?
2. What data modelling training have you had?
3. What data modelling experience have you had in your career?
4. What exposure have you had to the literature on data warehouse design? (e.g. Inmon, Kimball). Had you read or have you since read anything on data warehousing in any industry, computer journals, for example?

Existing database(s):

1. What existing systems was the data warehouse design based on.
2. How did you ensure that the information retrieved from the existing production database was useful (i.e. the information content was accurate)?
3. How reliable was the production system? Did you have to change the physical schema at all?
4. Was there an existing repository/data dictionary?

Existing business requirements:

1. Did you have any existing data models representing the business domain, or sub-groups of the business domain?
2. Was there a physical data model of the existing database(s)?
3. To what degree were the users involved in the original business requirements definition? (specifically the data requirements)

Data warehouse design:

1. Was a formal method used to construct the data warehouse? What were the main steps involved in this methodology?
2. Was a data model created as part of the data warehouse design? Who was involved in developing the data model?
3. When did the data modelling for the data warehouse commence?
4. What data modelling approach was followed? (For example, star schema, relational data modelling or both). What influenced this choice?
5. Can you provide a description of the method used to create the data model?
6. How often is the data model referenced? As far as you are aware, do the users understand the model?
7. Was a tool used to create the model (e.g ABC flow charter, CASE tool)?
8. If changes are made to the physical structure of the data warehouse is the data model updated to reflect this?

Data model validation:

1. What method was used to validate the data model? Who was involved in this validation process?
2. Were the end users involved in the validation process? What method was used to explain the data model(s) to the end users? How difficult did the end users find this validation process?
3. How far do you feel it is possible to guarantee the semantic integrity of the model?
4. Do you feel that the validation process was adequate?
5. Did the data warehouse project result in, or influence change on, the data that is collected operationally?

Project Outcome:

1. How successful has the design proven to be? i.e. how easy might it be to modify?
2. What kind of questions do you ask of the data cubes?

3. How many people have access to the data warehouse?
4. Is the data easy to access and understand?
5. Since you've been running the data warehouse has it been generally good? How fast is the query response time?
6. Do you have confidence in the quality of the data in the warehouse?
7. What would you do differently if you could re-design the data warehouse?
8. Has going through the activity of building the warehouse altered in any way the way the business is run?

Appendix 5

Critical Appraisal Guidelines for Single Case Study Research

CONDUCT OF RESEARCH

Research Method	1	2	3	4	5	6	7
1. Is a credible argument given for why a case study is appropriate?							
2. Are the philosophical stance and perspective of the authors stated?							
3. Is there evidence that any such bias is taken into account when performing data analysis?							
4. Is the history and context of the research clearly described?							
Transparency Of Process							
5. Are the aims and objectives of the study clearly stated?							
6. Are the criteria used to select the appropriate case and participants clearly described?							
7. Are approaches and techniques (and the rationale for their selection) for data collection and analysis stated clearly?							

Credibility of the Research								
8.	Does the study define and use some form of quality control measures?							
9.	Does the study describe an orderly process for the collection of data?							
10.	Does the study describe and employ a systematic way to analyse the data?							
11.	Does the study use appropriate theory to support the findings.							
12.	Does the study adequately describe how the conclusions were arrived at and how they are justified by the results?							
13.	Are assertions / conclusions made about data logical and coherent?							
14.	Are findings and conclusions grounded in the data?							
15.	Are data analysis and research findings confirmable (or have they been confirmed) by an outside expert?							
16.	Does the study place the findings in the context of IS practice?							
17.	Does the study place the findings in the context of IS research?							
18.	Acknowledge and describe any limitations to the study.							

Transferability of the Research								
19.	Are descriptions of setting, process and outcomes sufficiently rich to aid the judgments and decisions of other researchers regarding the transferability of the research to other contexts?							
20.	Does the study suggest if and how the findings might be transferable to other settings?							
21.	Have any opportunities for various forms of triangulation been exploited?							
Dependability of the Research								
22.	Is the research process auditable?							
23.	Is the research process open to scrutiny?							
24.	Are the bases for decision making and assertions / claims explicit?							
CONCEPTUAL SIGNIFICANCE								
25.	Does the study provide a clearly formulated question that describes an important IS issue or problem of interest							
26.	Has significant literature in the area of interest been accessed, supporting the selection of an appropriate theoretical framework to guide the research?							

27. Does the study lead to questions or issues for future research?							
PRACTICAL SIGNIFICANCE							
28. Could this research potentially make a helpful contribution to the work of practitioners in the field of IS?							
29. Does the research provide new insights into some aspect of IS discipline work?							
PRESENTATION OF RESEARCH							
30. Is the action research presented in such a way that there is evidence of logical rigour throughout the study?							
31. Are the links evident between a problem in the IS field, the literature review, theoretical framework, research method and design, and results / outcomes?							
32. Has the consumer of the research been identified?							
33. Is the research presented in an appropriate form and style to suit the consumer's objectives?							

Appendix 6
Q.S.R. NUD*IST4
Index Structure

Node Address	Node Name
(1)	/Base Data
(1 1)	/Base Data/Interviewee's
(1 1 2)	/Base Data/Interviewee's/Job Title
(1 1 3)	/Base Data/Interviewee's/Length of Time Employed
(1 1 4)	/Base Data/Interviewee's/Qualifications
(1 1 5)	/Base Data/Interviewee's/Data Modelling Experience
(1 1 6)	/Base Data/Interviewee's/Role on the MIS project
(1 1 7)	/Base Data/Interviewee's/Literature knowledge
(2)	/Levels of Meaning
(2 1)	/Levels of Meaning/Understanding
(2 1 1)	/Levels of Meaning/Understanding/Degree of Understanding
(2 1 2)	/Levels of Meaning/Understanding/Inhibiting factors
(2 1 2 1)	/Levels of Meaning/Understanding/Inhibiting factors/Complexity of software
(2 1 2 2)	/Levels of Meaning/Understanding/Inhibiting factors/Reports are limiting
(2 1 2 3)	/Levels of Meaning/Understanding/Inhibiting factors/Pre-defined cubes
(2 1 2 5)	/Levels of Meaning/Understanding/Inhibiting factors/Comprehensiveness
(2 1 2 6)	/Levels of Meaning/Understanding/Inhibiting factors/Complexity of AMS
(2 1 2 8)	/Levels of Meaning/Understanding/Inhibiting factors/Training inadequate
(2 1 2 9)	/Levels of Meaning/Understanding/Inhibiting factors/Teething problems
(2 1 2 11)	/Levels of Meaning/Understanding/Inhibiting factors/MIS resource
(2 1 2 12)	/Levels of Meaning/Understanding/Inhibiting factors/Aptitude & motivation
(2 1 2 14)	/Levels of Meaning/Understanding/Inhibiting factors/Data complexity
(2 1 2 14 1)	/Levels of Meaning/Understanding/Inhibiting factors/Data complexity/Structured codes
(2 1 2 14 2)	/Levels of Meaning/Understanding/Inhibiting factors/Data complexity/Application typing problem
(2 1 2 14 3)	/Levels of Meaning/Understanding/Inhibiting factors/Data complexity/Format of the data
(2 1 2 14 4)	/Levels of Meaning/Understanding/Inhibiting factors/Data complexity/Query design
(2 1 3)	/Levels of Meaning/Understanding/Validity claims
(2 1 3 1)	/Levels of Meaning/Understanding/Validity claims/Comprehensibility
(2 2)	/Levels of Meaning/Connotation
(2 2 1)	/Levels of Meaning/Connotation/Differentiated meaning between user groups
(2 2 2)	/Levels of Meaning/Connotation/Inhibiting factors

Node Address	Node Name
(2 2 2 1)	/Levels of Meaning/Connotation/Inhibiting factors/Cultural Issues
(2 2 2 3)	/Levels of Meaning/Connotation/Inhibiting factors/Dissidence
(2 2 2 4)	/Levels of Meaning/Connotation/Inhibiting factors/Branch knowledge of big pic low
(2 2 3)	/Levels of Meaning/Connotation/Validity claims
(2 2 3 1)	/Levels of Meaning/Connotation/Validity claims/Correctness
(2 2 3 1 4)	/Levels of Meaning/Connotation/Validity claims/Correctness/Data not live
(2 2 3 2)	/Levels of Meaning/Connotation/Validity claims/Ambiguity
(2 2 4)	/Levels of Meaning/Connotation/Consistency Issues
(2 3)	/Levels of Meaning/Intention
(2 3 1)	/Levels of Meaning/Intention/(for the generation of meaning)
(2 3 1 1)	/Levels of Meaning/Intention/(for the generation of meaning)/Inhibiting factors
(2 3 1 1 1)	/Levels of Meaning/Intention/(for the generation of meaning)/Inhibiting factors/Different reporting requirements
(2 3 1 1 2)	/Levels of Meaning/Intention/(for the generation of meaning)/Inhibiting factors/Reliance on one person
(2 3 1 1 3)	/Levels of Meaning/Intention/(for the generation of meaning)/Inhibiting factors/Response time
(2 3 1 1 4)	/Levels of Meaning/Intention/(for the generation of meaning)/Inhibiting factors/Data not compelling
(2 3 1 2)	/Levels of Meaning/Intention/(for the generation of meaning)/Validity claims
(2 3 1 2 3)	/Levels of Meaning/Intention/(for the generation of meaning)/Validity claims/Reliability
(2 3 1 2 7)	/Levels of Meaning/Intention/(for the generation of meaning)/Validity claims/Distrust
(2 3 1 3)	/Levels of Meaning/Intention/(for the generation of meaning)/Need is not there
(2 3 2)	/Levels of Meaning/Intention/(For the production of a data model from meaning)
(2 3 2 1)	/Levels of Meaning/Intention/(For the production of a data model from meaning)/Inhibiting factors
(2 3 2 1 1)	/Levels of Meaning/Intention/(For the production of a data model from meaning)/Inhibiting factors/Poorly thought out initial designs.
(2 3 2 1 3)	/Levels of Meaning/Intention/(For the production of a data model from meaning)/Inhibiting factors/Understanding the process and policies
(2 3 2 2)	/Levels of Meaning/Intention/(For the production of a data model from meaning)/MIS intention
(2 4)	/Levels of Meaning/Generation
(2 4 1)	/Levels of Meaning/Generation/Inhibiting factors
(2 4 1 1)	/Levels of Meaning/Generation/Inhibiting factors/Analysis specification outdated
(2 4 1 2)	/Levels of Meaning/Generation/Inhibiting factors/Lack of user involvement
(2 4 1 3)	/Levels of Meaning/Generation/Inhibiting factors/Limited logical design
(2 4 1 4)	/Levels of Meaning/Generation/Inhibiting factors/Constrained by AMS
(2 4 1 5)	/Levels of Meaning/Generation/Inhibiting factors/Complication of the data

Node Address	Node Name
(2 4 1 6)	/Levels of Meaning/Generation/Inhibiting factors/Project Timing
(2 4 2)	/Levels of Meaning/Generation/Validity Claims
(2 4 2 1)	/Levels of Meaning/Generation/Validity Claims/Rightness
(2 4 2 1 3)	/Levels of Meaning/Generation/Validity Claims/Rightness/Varying confidence in the data
(2 4 2 2)	/Levels of Meaning/Generation/Validity Claims/Truth
(2 4 2 2 9)	/Levels of Meaning/Generation/Validity Claims/Truth/DM validation
(2 4 2 2 9 8)	/Levels of Meaning/Generation/Validity Claims/Truth/DM validation/No motivation to validate
(2 4 2 2 9 9)	/Levels of Meaning/Generation/Validity Claims/Truth/DM validation/No validation
(2 4 2 2 9 10)	/Levels of Meaning/Generation/Validity Claims/Truth/DM validation/Lack of end user involvement
(2 4 2 3)	/Levels of Meaning/Generation/Validity Claims/Effectiveness
(2 4 2 3 1)	/Levels of Meaning/Generation/Validity Claims/Effectiveness/Difficult to modify
(2 4 2 3 2)	/Levels of Meaning/Generation/Validity Claims/Effectiveness/Denormalisation
(2 5)	/Levels of Meaning/Action
(2 5 1)	/Levels of Meaning/Action/Types of use or action
(2 5 2)	/Levels of Meaning/Action/What they would like to do
(2 5 3)	/Levels of Meaning/Action/Inhibiting factors
(2 5 3 1)	/Levels of Meaning/Action/Inhibiting factors/Providing info to the end users
(2 5 3 2)	/Levels of Meaning/Action/Inhibiting factors/No of users
(2 5 3 3)	/Levels of Meaning/Action/Inhibiting factors/Reliance on manual methods
(3)	/AMS
(3 1)	/AMS/Data input
(3 2)	/AMS/Design problems
(3 3)	/AMS/Poor database design
(3 4)	/AMS/Records unstable
(3 5)	/AMS/Lack of change control on AMS
(3 6)	/AMS/Anomalies in AMS