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

List of Figures

List of Tables

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.