Understanding Data Quality in Data Warehousing: A Semiotic Approach

Graeme Shanks and Peta Darke
Department of Information Systems
Monash University,
Melbourne, Australia
Email: graeme.shanks@sims.monash.edu.au

Abstract

The quality of data is a key factor in the success of data warehousing and there is strong evidence that data quality problems are becoming increasingly prevalent in practice. Much existing work on data quality focuses on the intrinsic quality of data in databases and provides lists of desirable data quality dimensions. This paper describes a comprehensive framework for understanding data quality that is based in semiotic theory. The framework is used to understand and analyse the data quality practices of a large Australian organisation within a data warehousing project. A number of potential uses of the framework in both data quality practice and research are identified.

1. Introduction

There is strong evidence that data quality problems are becoming increasingly prevalent in practice (Redman 1998, Wand and Wang 1996) and have significant social and economic impacts (Strong *et al.* 1997). This trend is particularly important for data warehousing where data is obtained from multiple sources and used by people who are far removed from the original data collection and may have little understanding of the nuances regarding the meanings of data items (Tayi and Ballou 1998). Much of the existing work on data quality focuses on the intrinsic quality of data in databases and provides lists of desirable data quality dimensions. These lists are often vaguely defined and not based in sound underlying theory. Although Wand and Wang (1996) provide definitions of intrinsic data quality dimensions that are anchored in ontological foundations (Weber 1997), there is a strong need for a comprehensive, theoretically-based data quality framework which covers both intrinsic and extrinsic data quality dimensions.

Data warehousing is an important contemporary issue in information systems practice. A data warehouse is a database intended to provide the data infrastructure for management support systems, including executive information systems and decision support systems (Shanks *et al.* 1997). An important motivation for data warehousing is to improve the process and outcomes of decision making within an organisation. Data quality is a key factor in the success of data warehousing (Kimball 1996, Wang 1998). Clearly, it is critical to organisations that data quality is understood and that procedures are in place to assure the quality of data in data warehouses.

In this paper we define a framework for understanding data quality which includes both intrinsic and extrinsic data quality dimensions. The framework is based in semiotic theory, the study of the use of signs and symbols to convey knowledge (Stamper 1992). A key feature of the framework is the separation of data quality goals from the means to achieve them. Other components of the framework include improvement strategies, measures, weightings and ratings. The framework provides researchers and practitioners with a sound, theoretically-based framework which will support further work in the development of data quality guidelines, evaluation procedures and empirical studies of data quality in practice.

The paper first defines data quality, and critically reviews existing work in data quality. The next section of the paper describes the philosophical and theoretical foundations on which the framework is based. The fourth section of the paper describes the framework and its components. The following section describes how concepts in the framework have been used in a case study to understand data quality within a data warehousing project. The paper concludes with a discussion of the potential use of the framework in both data quality practice and research.

2. Data Quality in Data Warehousing

A data warehouse is a "subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decisions (Inmon and Hackathorn, 1994). *Subject-oriented* and *integrated* indicates that the data warehouse is designed to span the functional boundaries of legacy systems in order to provide management with an integrated view of the data. *Time-variant* concerns the historical or time-series nature of the data in a data warehouse, which enables trends to be analysed. *Non-volatile* indicates that the data warehouse is not continuously updated like an OLTP (on-line, transaction processing) database. Rather, it is uploaded periodically, say daily, with data from internal and external data sources. The data warehouse is specifically designed for retrieval rather than for update integrity and transaction throughput. The idea of data warehousing is not new, it has been one of the aims of management support since the 1960s (Martin, 1982). Although the goal proved to be elusive, recent developments in technology have prompted many current data warehousing projects.

Quality is defined as "fitness for purpose" and should include not only the intrinsic characteristics of the data itself but also assessments of users of the data (Strong *et al.* 1997). This means that data in the data warehouse must be usable and useful to the consumers of the data and support their effective work practices.

Two different levels of data quality may be defined within data: structure (metadata) and content (data). The quality of the structure of a data warehouse is defined as the quality of the conceptual model that is the basis for the design of the data warehouse. Quality metadata is important for all stakeholders in the data warehousing process so that they understand what the data warehouse contains and how to access data in the data warehouse. Quality data is important so that users of the data warehouse can readily assess and understand data in the data warehouse and use the data effectively in their decision making activities. This paper focuses on content (data) quality.

Much of the work in data quality consists of lists of desirable data quality dimensions (see Wand and Wang (1996: p92) for a useful summary of data quality dimensions in the literature). These lists typically include dimensions such as accuracy, reliability, importance, consistency, precision, timeliness, understandability, conciseness and usefulness. These dimensions are often overlapping, vaguely defined, ambiguous and not soundly based in theory.

Some frameworks have been developed which organise and structure important concepts in data quality. Wang and Strong (1996) organise data quality dimensions into four categories: intrinsic, contextual, accessibility, and representational. The intrinsic category contains dimensions that define the quality of data in its own right. The contextual category contains dimensions that define data quality within the context of the task at hand. The accessibility and representational categories contain dimensions that define data quality in terms of the information systems, which store and deliver the data. This categorisation is useful as it explicitly highlights the separation of intrinsic data qualities from other categories of data quality. Kahn et al. (1997)

developed a framework based on the total data quality management literature and categorise data quality dimensions as either sound, useful, usable or effective. This framework is potentially useful in that it emphasises both the product and service nature of data quality and provides a means of determining the level of sophistication of data quality management within organisations. However, neither of these frameworks provides a set of data quality dimensions that are soundly based in theory.

Wand and Wang (1996) have developed a framework that anchors data quality dimensions in the ontological foundations defined by Bunge (see Weber (1997) for a detailed discussion of Bunge's ontology applied to information systems modelling). The framework provides formal definitions for several data quality dimensions but is limited to intrinsic (or system oriented) data quality dimensions. Wand *et al.* (1995) suggest the use of semiotics as an alternative theoretical basis to Bunge's ontology.

3. Theoretical Foundations of the Framework

In defining the theoretical foundations of our framework we first establish the philosophical position we adopt and then we use semiotic theory to define various levels of the use of symbols as a means of conveying knowledge. The philosophical position is based on Mingers (1995) and leads to definitions for data, information and meaning. The semiotic levels will lead to the definition of quality goals.

3.1 Philosophical Position

We adopt a *realist* ontological position and accept that the physical world consists of things with attributes that are related in a causal way (Burrell and Morgan 1979). The real world system is described in terms of its states (the values of its attributes at a certain point in time), events and laws (that determine the allowed combinations of values of attributes) (Wand and Wang 1996). In an information system, the state of the real world system is represented by symbols (textual, numeric, or graphical). These symbols may be used only when there is an agreed set of rules governing their interpretation by people: a system of connotation. Groups of people within organisations need to share the same system of connotation in order to effectively communicate with a set of symbols (Mingers 1995). For example, a set of codes used to represent customer types in different databases must follow the same set of rules to permit comparison or consolidation and use by people in different parts of the organisation.

Data is a collection of symbols that are brought together because they are considered relevant to some purposeful activity (Mingers 1995). For example, a set of symbols can be brought together to represent a customer type code, customer type description and default credit limit for the purpose of facilitating sales of goods.

Information is carried by symbols and is an objective (although abstract) commodity that exists independently of any person who may interpret the symbols. The information carried by a symbol is causally implied by the occurrence of the symbol. For example, the symbol "zero" for default credit limit implies that cash or cheque is required for payment, and that the customer may have defaulted on payments in the past. The information carried by a symbol relates to who produces it, why and how it was produced and its relationship to the real world system state it signifies. Information is the propositional content of a symbol: that which is implied by the occurrence of the sign (Mingers 1995).

Although a realist ontological position is adopted, a *subjectivist* epistemological position is assumed. That is, our understanding of the world depends on our prior knowledge and experience. Not all the information contained in a data set is available to people who interpret it. For example, the symbol "zero" may have been allocated to a new customer for default credit limit. This may be standard practice until a credit check can be carried out. However, a user of the data may not have this knowledge and may interpret the data as indicating a high risk customer who has probably defaulted on credit previously.

Meaning is defined as the particular meaning that people derive from symbols and is generated from the information that accompanies data. Meaning is inherently intersubjective as it depends on common systems of connotations of the producer and interpreter of the data. Meaning involves a person understanding a symbol, relating that understanding to their knowledge and experience, and then forming intentions for action based on their understanding (Mingers 1995).

The definitions of data, information and meaning correspond to the everyday usage of the terms (Mingers 1995). A summary of the philosophical position and important definitions is provided in Table 1 below.

Table 1 Summary of Philosophical Position and Important Definitions

Ontological Position	*Realist* - the real world system consists of things and attributes that are related in a causal way, and is described in terms of its states (the values of its attributes at a certain point in time), events and laws (the allowed combinations of values of attributes)
Epistemological Position	*Subjectivist* - our understanding of the world depends on our prior knowledge and experience.
Data	Data is a collection of symbols which signify real world system states and are brought together because they are considered relevant to some purposeful activity
Information	Information is an objective commodity carried by symbols and relates to who produced it, why and how it was produced and its relationship to the real world state it signifies
Meaning	Meaning is inherently intersubjective and is generated from information by a person understanding and interpreting data.

3.2 Semiotic Theory

Semiotic theory involves the use of symbols to convey knowledge. It has a long history dating from the ancient Greek philosophers through John Locke in the seventeenth century to Peirce and Saussure in the twentieth century (Stamper 1992). Semiotics uses the sign or symbol as its primitive notion and for analytical purposes has traditionally been divided into several levels. Stamper (1992) describes six semiotic levels: physical, empirical, syntactic, semantic, pragmatic and social. The physical and empirical levels concern the physical media and use of the physical media for communication of symbols and will not be considered in this paper. The social level concerns the purpose of information in relation to social norms and social change and will also not be considered in this paper.

The three semiotic levels that are of interest in discussing data quality are the syntactic, semantic, and pragmatic levels. Although these levels are separated for analytical convenience, they are closely interrelated and build on each other. The three levels are briefly discussed below. The relationships between the semiotic levels and data quality are shown in Figure 1 below.

Syntactic — Concerns the *structure* of symbols and focuses on form rather than content. The syntax consists of the valid syntactic categories and the rules that govern their form. If the syntax is formally defined then symbolic forms may be transformed into other symbolic forms. Two symbolic forms are equivalent if they may be transformed into one another.

Semantic — Concerns the *meaning* of symbols. Meanings are assigned to data by people depending on their prior knowledge and experience. Meanings are intersubjective and continuously constructed. The mapping of a symbol to a real world state is possible but may be different for different people. The semantic *meaning* is built upon the syntactic *structure*.

Pragmatic — Concerns the *usage* of symbols and is dependent on the task of the person using the data. Intention for action (either psychological or physical) is imputed on a symbol by its creator and its interpreter (Stamper 1992). The pragmatic understanding of a symbol is dependent on the social context in which the symbols have their effect, and can only be acquired in a specific situation. The pragmatic *usage* is built upon the semantic *meaning*.

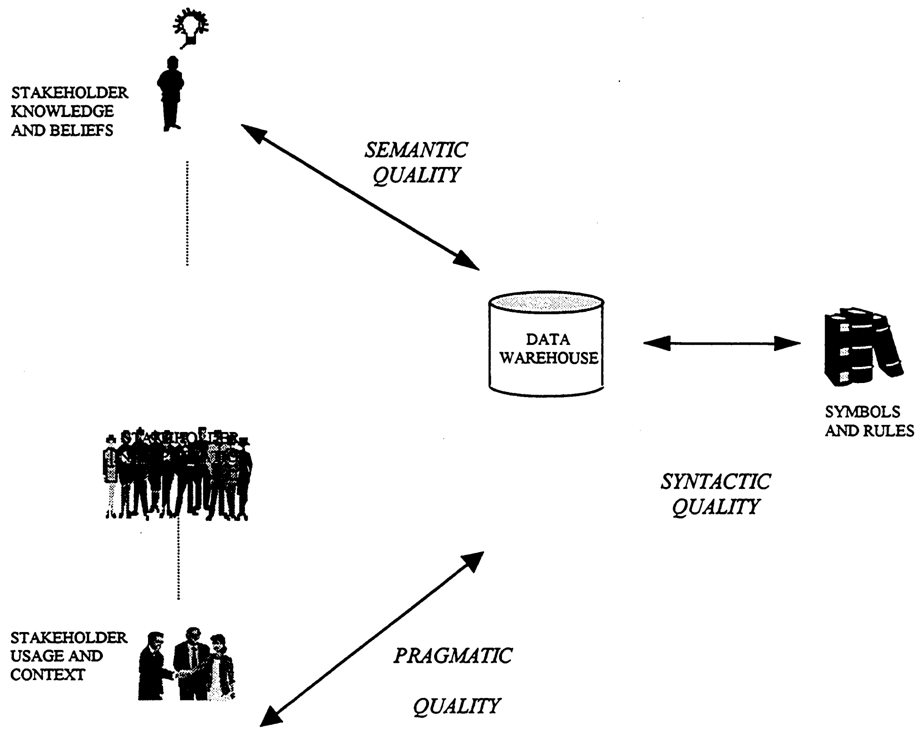

Figure 1 Semiotic Levels in Understanding Data Quality in a Data Warehouse

4. A Framework for Understanding Data Quality in a Data Warehouse

The framework for understanding data quality in a data warehouse consists of a number of components and their interrelationships. The key component of the framework is a set of data quality goals. These goals are defined for each of the three semiotic levels. The means to achieve each of the goals are then defined in terms of desirable data quality properties and related improvement strategies. To support evaluation of data quality, measures for each property are defined. The final component of the framework is stakeholders: those responsible for producing, maintaining, and consuming data. Figure 2 shows the components of the framework and their interrelationships.

4.1 Syntactic Data Quality

Syntactic data quality concerns the *structure* of data. The goal for syntactic data quality is *consistency* where data values for particular data elements in the data warehouse use a consistent symbolic representation (Ballou et al. 1996, Wang et al. 1995). This may be within a single data file where all data values should conform to the same strict data type definitions, or between data files where values need to be consistent to permit consolidation and comparison (Mattison 1996). Consistency is particularly important in coding schemes throughout an organisation, for example customer codes, part codes and region codes.

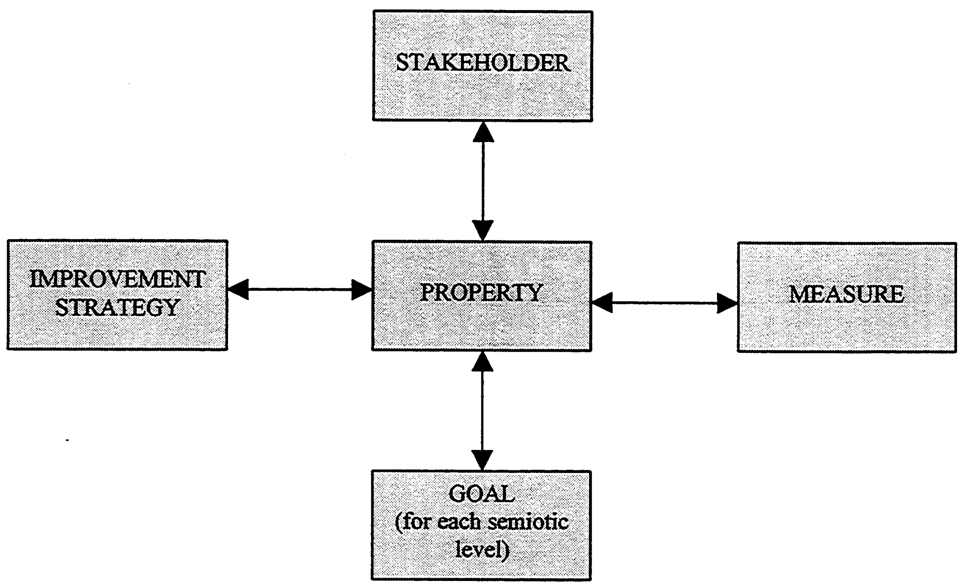

Figure 2 Framework for Understanding Data Quality in a Data Warehouse

The means to ensure consistency is to have a well-defined, perhaps formal, syntax for all data elements. The improvement strategies include the development of a corporate data model, with syntax rules for data elements having an enterprise-wide perspective This is particularly important in a data warehousing environment where common syntax rules are essential for sourcing, cleansing and loading data from legacy systems (Inmon and Hackathorn 1994). A second improvement strategy is to have automatic syntax checking at data entry or to have human data producers well trained in the syntax rules. A measure for consistency is to express the ratio of the number of inconsistent data values to the total number of data values for each data element in the data warehouse.

4.2 Semantic Data Quality

Semantic data quality concerns the *meaning* of data. The goals for semantic quality are *comprehensiveness* and *accuracy* (Tayi and Ballou 1998, Wang *et al.* 1995). Comprehensiveness is concerned with the extent to which for each relevant state in the real world system there is a data value in the data warehouse. Accuracy is concerned with how well the data values in the data warehouse correspond to the state of the real world. As each stakeholder may have different prior knowledge and experience, different stakeholders may have different views on the comprehensiveness and accuracy of the data warehouse.

The properties necessary to ensure comprehensiveness and accuracy are consistency (the syntactic data quality goal) and the data quality dimensions defined by Wand and Wang (1996). These dimensions are defined by analysing mappings between symbols in the data warehouse and stakeholder understanding of real world structures and events. The dimensions are: complete, unambiguous, meaningful and correct. For data in the data warehouse to be complete, the mapping between the real world system and data values in the data warehouse must be exhaustive (i.e. each state in the real world system is mapped to a data value in the data warehouse). For data to be unambiguous, no two states of the real world system should be mapped into the same data value of the data warehouse. For data to be meaningful there should be no data values in the data warehouse which cannot be mapped to a real world system state. For data to be correct, real world system states must not be mapped onto the wrong data values in the data warehouse (see Wand and Wang (1996) for detailed definitions and discussion).

The improvement strategies to achieve a comprehensive and accurate data warehouse include training the producers of data in the importance of comprehensive and accurate data. Another important strategy is to minimise the number of data transformations and data transcriptions from when the data is first captured until it is stored in the data warehouse (Haebich 1996). Measures for completeness, ambiguity, meaningfulness and correctness include surveying population samples and comparing data values in the data warehouse with real world system states.

4.3 Pragmatic Data Quality

Pragmatic data quality concerns the *usage* of data. The goals for pragmatic quality are *usability* and *usefulness* (Kahn *et al.* 1997). Usability is the degree to which each stakeholder is able to effectively access and use the data in the data warehouse. Usefulness is the degree to which the data supports the stakeholder in accomplishing their tasks within the social context of the organisation. Usability and usefulness will vary between different stakeholders due to their different interpretations of the meaning of data values and the different nature of their tasks.

The properties necessary to ensure usability and usefulness include consistency, completeness and accuracy (the syntactic and semantic data quality goals), timeliness, ease of understanding, accessibility, conciseness, and reputation. Timeliness is the extent to which the information is up-to-date for the task at hand. Ease of understanding refers to the extent to which data in the data warehouse is understood by stakeholders. Accessibility is the ease with which data can be retrieved and manipulated. Reputation is the extent to which data is highly regarded in terms of its source, content and credibility (Kahn *et al.* 1997).

The improvement strategies to achieve a usable and useful data warehouse include monitoring data consumers' use of the data warehouse to ensure data is up-to-date for the task at hand, and using explanation and visualisation techniques to assist stakeholder understanding of data (Shanks and Darke 1998). High quality data delivery systems are essential to provide appropriate access for data consumers. Data tagging is a useful way of providing data quality information to data consumers (Chengalur-Smith *et al.* 1997). Measures for usability and usefulness are mostly highly subjective and include length of time from last update (for time sensitive data), surveys of stakeholder beliefs about data in the data warehouse (subjective stakeholder rating) and the impact on their decision making processes and outcomes.

Table 2 Summary of Goals, Properties, Improvement Strategies and Measures

Semiotic Level	Goal	Property	Improvement Strategy	Measure
Syntactic	Consistent	Well-defined (perhaps formal) syntax	Corporate data model, Syntax checking, Training for data producers	Percentage of inconsistent data values
Semantic	Compreh-ensive and Accurate	Complete, Unambiguous, Meaningful, Correct	Training for data producers, Minimise data transformations and transcriptions	Percentage of errors in data or population sample
Pragmatic	Usable and Useful	Timely, Understood, Concise, Easily Accessed, Reputable	Monitoring data consumers, Explanation and visualisation, High quality data delivery systems, Data tagging	Time of update, User surveys, Effect on decision making processes and outcomes

4. 5 Stakeholders, Weightings, and Ratings

Four types of stakeholder are defined in the framework: data producers, data custodians, data consumers, and data managers (Strong et al. 1997, Wang 1998). Data producers are those who create or collect data for the data warehouse. Data custodians are those who design, develop and operate the data warehouse. Data consumers are those who use the data in their work activities. Data managers are those who are responsible for managing the entire data warehousing process.

Two further components are introduced into the framework to enable individual stakeholders or stakeholder groups to evaluate data quality or identify data quality requirements in practice. These two components are weightings, which are allocated to properties to indicate their relative importance to the task at hand, and ratings (or scores) which are assigned using measures. Weightings and ratings can be applied at various levels of data granularity from database file, data attribute or data value. Figure 3 below shows how these additional components extend one section of the framework. From the figure it can be seen that weightings are allocated by a stakeholder to a property for a certain data element and ratings are assigned by a stakeholder to particular data elements using specific measures.

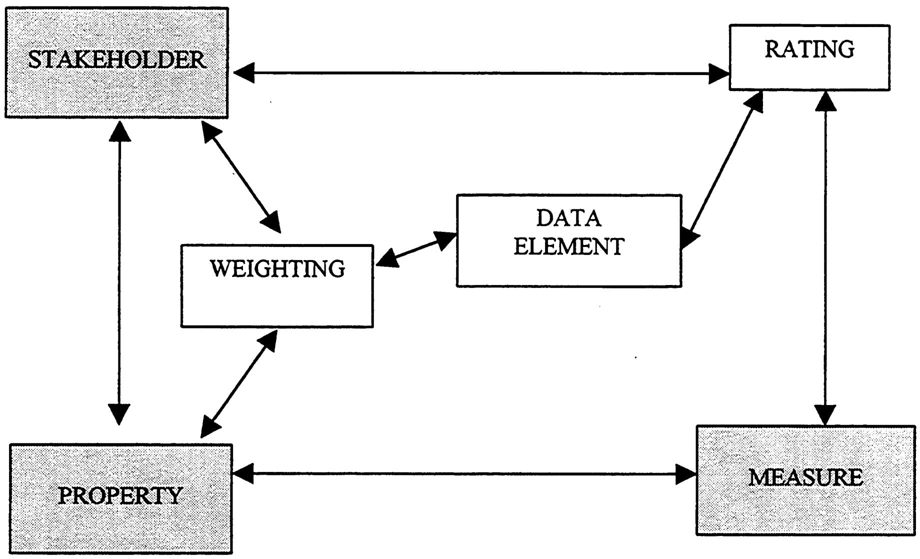

Figure 3 Part of the Framework including Weightings and Ratings

5. Applying the Framework in Practice: a Case Study

This section of the paper describes a case study of the data quality practices of a large Australian organisation within a data warehousing project. The case study was conducted in 1997. The provides a range of financial products and services to its members. These include vehicle, home and life insurance as well as personal loans. The organisation also provides travel and accommodation plans and bookings for its members. The organisation currently has 1.3 million members.

5.1 The Data Warehousing Project

The organisation had decided to implement a data warehouse in order to provide better customer information and to improve customer services. The interest in data quality was primarily motivated by the direct customer marketing function. The number of incorrectly addressed mailouts needed to be reduced and improved demographic information was expected to lead to better targeted mailouts for particular products and services.

A data quality committee was formed consisting of eight business personnel and two information technology personnel. At the time of data collection, the committee had existed for six months and was still developing policies and procedures and dealing with data ownership and data sharing issues. The main purpose of the committee was to gain high level understanding and sponsorship for data quality within the organisation.

Data quality was understood within the organisation as being a multi-dimensional concept. The various dimensions recognised as important to the organisation were correctness, completeness, consistency and integrity. However, these dimensions sometimes had different meanings in different sections of the organisation. For example, the concept of "customer" meant "member" to those concerned with mailouts of the monthly magazine. For those concerned with marketing financial products and services, it meant "any person or organisation that interacted in any way with the organisation". In this way the customer data may be complete for one stakeholder and incomplete for another.

5.2 Data Quality Initiatives

A corporate data model had been developed to standardise data definitions across the organisation. Historically, membership codes had been kept consistent throughout the organisation's databases, so consolidation of member data for the data warehouse was simplified. However, some transformation of data formats was required in consolidating data from various legacy systems into the data warehouse.

Of more concern to the organisation were the problems of duplicate customers, inaccurate customer address details, inaccurate dates for customer business transactions and referential integrity errors between data files. A particular problem was the difference between the customer and member concepts. Members pay an annual subscription and are allocated member numbers. However, a spouse of a member may take out an insurance policy and is therefore a customer but without a member number.

The organisation regularly conducts random sampling of customer data by contacting customers by telephone and checking the data values in the data warehouse. At other times the organisation instructs its telephone operators to check the details of all customers who ring with inquiries. Statistical records are retained of the results of these data quality checks. These are then used to estimate the cost to the organisation of data quality problems in terms of incorrect mailouts and lost opportunities.

The main focus of the organisation in addressing data quality problems is to fix the problem at the data source and reduce the number of phases of data transformation and transfer. Most of the effort directed at improving data quality to date has been aimed at reducing the number of incorrectly addressed customer mailouts and returned customer mailouts. In the future more effort will be directed at including demographic data and sales data in the data warehouse to facilitate more sophisticated customer analysis.

5.3 Analysis of the Case Study using the Framework

The three levels within our data quality framework provide a useful means to analyse the case study data. At the syntactic level, the goal of having a consistent symbolic representation for data was well recognised. The development of a corporate data model provided standardised data definitions for the organisation and facilitated the design of transformation procedures for data with different formats. Improvement strategies at the syntactic level may be automated, as human interpretation of data is not required. Consistency of representation is a basic prerequisite for loading data into a data warehouse. Syntactic data quality is perhaps the easiest of the quality goals to achieve and in the case study was the first to be addressed.

At the semantic level, the goal of having comprehensive and accurate data was recognised. For customer data, data quality at the semantic level was understood to be concerned with the mappings of data values in the data warehouse to the properties of people in the "real world system". Directly contacting random samples of customers will help the organisation understand the extent of its problems with meaningfulness and ambiguity of data. Including data quality checks when customers contact the organisation will help the organisation understand the extent of its problems with completeness and correctness of data. Improvement strategies at the semantic level require human action and interpretation of data to be effective. Achieving the goal of semantic data quality requires human interpretation and a higher level of sophistication than the goal of syntactic data quality. The meaning assigned to symbols depends on the knowledge and beliefs of stakeholders. The focus to date has been mostly on customer data and much more work at the semantic level will be required as the data warehouse is extended to include other data.

At the pragmatic level, the goal of having usable and useful data is less important than consistency, comprehensiveness and accuracy to the organisation at this point in time. The usage of the data under consideration is currently limited to mailouts that involve customer data. When decision-making tasks that involve data in the data warehouse become more complex, and involve data other than simple customer data, properties and improvement strategies at the pragmatic level are expected to be important. Pragmatic data quality improvement will require organisations to focus their attention on individual decision-makers, their decision making styles, their tasks and the organisational context within which the task occurs. As pragmatic data quality builds on syntactic and semantic data quality, it is unlikely that much effort will be expended on pragmatic data quality until syntactic and semantic data quality improvement procedures are well developed and more sophisticated uses of the data warehouse are deployed.

6. Conclusion

This paper has proposed a framework for understanding data quality in a data warehouse that is based in semiotic theory. The framework encompasses both intrinsic and extrinsic data quality, and separates data quality goals from the means to achieve them. It extends other approaches to data quality which provide lists of data quality dimensions that are frequently overlapping, vaguely defined, ambiguous, and not soundly based in theory.

The three semiotic levels focus attention on the structure, meaning and usage of data. The structure of data concerns its symbolic representation and is the simplest of the levels to understand. The meaning of data requires a clear statement of philosophical position to be made. A realist ontological position and a subjectivist epistemological position form the basis of the definition of "meaning" at the semantic level and permit the inclusion of Wand and Wang's (1996) ontologically anchored data quality dimensions. The meaning of data depends on human interpretation. The usage of data encompasses the stakeholder, the task and the context within which the task occurs. This is the most complex of the three levels.

Several research areas concerning both the framework itself and its application are currently being investigated. These include:

- Refinement of the definitions of data quality goals, properties and improvement strategies for each of the semiotic levels. More precise definitions are required to enable improved measures and better analysis of data quality dimensions from other sources;

- Identification of data quality tagging metadata information (for example, information that the data values in a particular field are on average 70% correct). Empirical studies of the effect of data quality tagging on decision making processes and outcomes are also planned;

- Further use of the framework as a means of analysing case study data from empirical studies of the practice of data quality management. This will help to develop an understanding of the level of sophistication of data quality management in organisations;

- Use of the framework components to develop methodologies for improving the process of data quality management within organisations;

- Development of automated tools to support the measurement of data quality in data warehouses. The tools should include weightings of quality dimensions by the various stakeholders involved in the data warehousing process and measures of quality for data elements. This information should support decisions about where to focus effort in data quality management.

8. References

Ballou, D. and Tayi, G. "Managerial Issues in Data Quality", Proceedings of the 1996 International Conference on Information Quality, MIT, Boston, (1996).

Burrell, G. and Morgan, G. Sociological Paradigms and Organisational Analysis, Heinemann, London, (1979).

Chengalur-Smith, I.N., Ballou, D. and Pazer, H.L. "The Impact of Data Quality Tagging on Decision Complacency", Proceedings of the 1997 Conference on Information Quality, Massachusetts Institute of Technology, (1997), pp. 209-226.

Haebich, W "A Methodology for Data Quality Management", Proceedings of 1st Australian DAMA Conference, Melbourne, (1996).

Inmon, W. and Hackathorn, R. Using the Data Warehouse, New York: John Wiley and Sons, (1994).

Kahn, B., Strong, D.M. and Wang, R.Y. "A Model for Delivering Quality Information as Product and Service", Proceedings of the 1997 Conference on Information Quality, Massachusetts Institute of Technology, (1997), pp. 80-94.

Kimball, R. The Data Warehousing Toolkit, John Wiley, (1996).

Mattison, R. Data Warehousing: Strategies, Technologies and Techniques, McGraw-Hill, (1996).

Mingers, J.C. "Information and meaning: foundations for an intersubjective account", Information Systems Journal, 5, (1995), pp. 285-306.

Redman, T. "The Impact of Poor Data Quality on the Typical Enterprise", Communications of the ACM, 41, 2 (1998), pp. 79-82.

Shanks, G., O'Donnell, P. and Arnott, D. "Data Warehousing: A Preliminary Field Study", Proc. 8th Australasian Conference on Information Systems, University of South Australia, Adelaide, (1997), pp. 350-365.

Shanks, G. and Darke, P. "Incorporating Context to Improve Understanding of a Data Warehouse", in Proc. IFIP Working Group 8.3 Conference on Decision Support Systems, Bled, Slovenia, (1998).

Stamper, R. "Signs, Organisations, Norms and Information Systems", Proc. 3rd Australian Conference on Information Systems, Wollongong, (1992).

Strong, D.M., Lee, Y.W. and Wang, R.Y. "Data Quality in Context", Communications of the ACM, 40, 5 (1997), pp. 103-110.

Tayi, G. and Ballou, D. "Examining Data Quality", Communications of the ACM, 41, 2 (1998), pp. 55-57.

Wand, Y., Monarchi, D.E., Parsons, J., Woo, C.C. "Theoretical Foundations for Conceptual Modelling in Information Systems Development", Decision Support Systems, 15 (1995), pp. 285-304.

Wang, Y., Reddy, M.P. and Kon, H.B. "Toward Data Quality: An Attribute-based Approach", Decision Support Systems, 13, 3-4 (1995), pp. 349-372.

Wang, Y. and Strong. D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers", Journal of Management Information Systems, 12, 4 (1996), pp. 5-34.

Wand, Y. and Wang, R. "Anchoring Data Quality Dimensions in Ontological Foundations", Communications of the ACM, 39, 11 (1996), pp. 86-95.

Wang, R.Y. "A Product Perspective on Total Data Quality Management", Communications of the ACM, 41, 2 (1998), pp. 58-65.