Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

An Exploratory Study Investigating Data Quality in the Healthcare Industry: What are the Implicatons for Daya Warehousing?

Michael S. Gendron
Central Connecticut State University, mgendron@gcstech.com

Marianne J. D'Onofrio
Central Connecticut State University, donofrio@ccsu.edu

An Exploratory Study Investigating Data Quality in the Healthcare Industry: What are the Implications for Data Warehousing?

Michael S. Gendron, Central Connecticut State University, mgendron@gcstech.com
Marianne J. D'Onofrio, Central Connecticut State University, Donofrio@ccsu.edu

ABSTRACT

Maintaining quality data and information challenges today's organizations, and is essential to good decision-making and competitive advantage. The purpose of this study is to understand the dimensions of data quality and to learn how managers perceive them. The 20 dimensions of data quality posited by Wang and Strong were used in this study. The relative importance of those dimensions by healthcare industry managers was assessed.

Study findings indicate that all 20 of Wang and Strong's data quality descriptive dimensions are important to the healthcare industry. Additionally, data from this exploratory research seems to suggest that there may not be a "one size fits all" model for data quality.

Keywords: Information, Data, Quality, Healthcare, Data-Warehousing

Category: Technical, Data Warehousing

INTRODUCTION

Data warehouses contain subject-oriented, integrated, time-variant, non-volatile collections of data in support of management decision-making processes to make timely, accurate decisions (Rudra & Yeo, 1999) Leaders in organizations in the 21st century are challenged to make decisions in a fast-paced, changing, and uncertain environment. The decisions made by corporate leaders reverberate throughout the corporation as the decisions are operationalized. Implementation strategies are developed and policies and procedures for various corporate activities result. For these activities to be of value to the business, the quality of the decisions and the quality of the data and information used in making the decisions leading to these activities are of paramount importance. (Ballou & Tayi, 1999). Thus, it becomes incumbent on today's organization to ensure that quality data and information are available and used in corporate decision-making.

Management at all levels of the organization relies on data and information and its flow throughout the organization for strategic, tactical, and operational decision-making. In many cases their main source of information is a decision support system or executive information system supported by a data warehouse. Those data warehouses are fed by the organization's online transaction processing (OLTP) systems, as well as other data sources, and are implemented to support decision-making processes within an organization. Current technology allows us to create large data warehouses to support business intelligence, but data quality is a concern (Celko & McDonald 1995, English, 1996). The increasing reliance on data warehouses as a critical resource in organizations requires that organizations manage this resource and improve its quality (Lee, 1999). In fact, substantial social and economic impacts result from poor data quality.(Wang & Strong, 1996)

The health care industry includes many types of organizations, which are data driven and in which quality of data is of paramount importance. Public health agencies, health maintenance organizations, and pharmaceutical companies are three of the more significant ones. Each of these organizations has data quality issues. For example, the effect of poor data quality in the public health sector can be seen in death certificate data (Altman, 1998). If that data is poor quality it can affect public health policy driven resource allocation. The follow-up of new pharmaceuticals, once the FDA permits them to go to market (Sharpe, 1998), also suffers from poor data quality. New pharmaceuticals are usually not monitored adequately once they are permitted to be marketed and any response to adverse outcomes is reactive, rather than proactive. Some health maintenance organizations' lack of good internal management information systems (Winslow, 1997) has caused their demise. These data quality issues are exacerbated by transferring data of less than optimal quality from OLTP system into data warehouses. The data are then readily available to many more people and are often used for purposes other than those for which the data was originally collected.

PURPOSE OF THE STUDY

By describing the variability in data quality dimensions and their importance rating among management levels in three healthcare related organizations, this research attempts to gain a better understanding of data quality in the healthcare milieu so we will be better equipped to combat both source data quality problems, as well data pollution that is found in data warehouses. This will be especially helpful in further research that attempts to create metadata standards within the healthcare industry.

THEORY BASE FOR RESEARCH

The descriptive work of Wang (1996) and his generic framework for defining data quality provides the theoretical underpinning for this research. Wang and Strong's article (1996), *Beyond accuracy: what data quality means to data consumers,* discusses the design and implementation of a study to develop "a framework that captures the aspects of data quality that are important to consumers."(Wang & Strong, 1996) Their framework posited four categories of data quality:

1. Intrinsic data quality denoting accuracy, objectivity, believability, and reputation.
2. Contextual data quality indicating relevancy, value-added, timeliness, completeness, and amount of data.
3. Representational data quality consisting of interpretability, ease of understanding, concise representation and consistent representation.
4. Accessibility data quality meaning access and security.

Wang and Strong's (1996) research resulted in the development of 20 dimensions of data quality.

This research extends the work of Wang and Strong (1996) by examining these 20 dimensions of data quality within a different population, healthcare professionals, and by investigating the variability among three sub-populations of healthcare professionals within different organizational levels. There are many data quality frameworks that have been put forth (see for example, Agmon & Ahituv 1987, King and Epstein 1989, Delone & McLean 1992, Wang and Strong, 1996, Firth 1997, Rudura & Yeo 1999). This research tests one of those frameworks to determine its strengths within the healthcare industry.

STUDY OBJECTIVES AND METHODOLOGY

The general objectives of this research were to:

1. Develop an understanding of the importance of the data quality dimensions trace-ability and cost-effectiveness within the healthcare industry.
2. Explore healthcare managers relative importance rating of the 20 dimensions of data quality posited by Wang.

The hypothesis testing was setup to determine if there is any significant variability within the healthcare industries perceptions of data quality.

A survey instrument was developed, using Wang's dimensions, requesting respondents to rate the importance of the 20 data quality dimensions to decisions they make on their job. For each item, respondents were presented with an ordinal scale mapped as follows: 1 - extremely important, 2 – very important, 3 – important, 4 – not very important, 5 – not important at all.

Respondents were from three sub-populations in the healthcare industry (pharmaceutical companies, health maintenance organizations, and public health agencies). The sampling frame was drawn from a national database of businesses maintained by *info*USA. COM. *Info*USA.com is a database provider that maintains lists of executives by industrial classification.

FINDINGS

The original work completed by Wang contained 20 dimensions of data quality, which he pared down to 15 dimensions and subsequently collapsed into 4 categories. Wang's data quality model seems to contain a good set of descriptive dimensions that permit us to understand data quality. Wang's work posited a 4-category model for data quality using these 15 dimensions not tied to any specific industry. However, data from this exploratory research seems to suggest that there may not be a "one size fits all" model for data quality. This is due to several reasons: 1) the 20 dimensions originally posited by Wang are all relevant to the healthcare milieu, since they all have been rated as important by healthcare managers, 2) when collapsing multiple dimensions into a single category, we might lose richness when we attempt to describe data, and 3) when creating metadata it would be simpler to use all 20 dimensions, rather than forcing individuals to use categories with underlying dimensions.

DISCUSSION

In light of the findings of this exploratory study, the dimension rankings of data quality may be associated with organization types and management levels. To ascertain the relationship among dimensions, organization types and management levels, further research might investigate the healthcare industry, utilizing a larger, more robust stratified random sample and investigate more demographic and background data about the organizations.

Various multi-dimensional data quality models can be found in the literature. This study supports the notion that data quality is a multi-dimensional concept, rather than a unitary one, and also suggests that data quality is domain specific and not a universal concept that transcends industries. This fact is equally important when building organizational schemas for data warehouses and when creating metadata so users can assess data quality.

BIBLIOGRAPHY

Altman, L. (1998, December 22). The doctor's world: getting it right on the facts of death. The New York Times, pp. Health and Fitness.

Agmon, N., & Ahituv, N. (1987). Assessing data reliability in an information system. Journal of Management Information Systems, 4(2), 34-44.

Ballou, D. P., & Tayi, G. K. (1999). Enhancing data quality in data warehouse environments. Communications of the ACM, 42(1), 73-78.

Celko, J., & McDonald, J. (1995). Don't warehouse dirty data. Datamation, 41(19), 42-53.

Delone, W.H. and McLean, E.R. (1992) Information Systems Success: The quest for the Dependent Variable, *Information Systems Research*, 3(1), 60-95.

English, Larry P. (1996). Help for Data Quality Problems - A number of automated tools can ease data cleaning and help improve data quality. INFORMATIONWEEK, 600.

King, W. R., & Epstien, B. J. (1983). Assessing information system value: an experimental study. Decision Sciences, 14, 34-45.

Lee, Y. W. (1999). Why is "know-why" knowledge useful for solving information quality problems? Available: http://hsb.baylor.edu/ramsower/ais.ac.96/papers/LEE3.htm.

Rudura, Amit & Emile Yeo. (1999) Working Paper Series: Achieving Data Quality and Consistency in Data Warehousing Among Some Large Organisations in Australia - Key Issues. School of Information Systems, Curtin University of technology, 9907. 1-18.

Sharpe, R. (1998, June 24). FDA drug monitoring system under fire. The Wall Street Journal, pp. B5.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-34.

Winslow, O. (1997, December 11). Oxford health plans nemesis - data quality. The Wall Street Journal, pp. B1.