DATA QUALITY: HOW THE FLOW OF DATA INFLUENCES DATA QUALITY IN A SMALL TO MEDIUM MEDICAL PRACTICE

Marlon Parker, Cornell Stofberg, Retha De la Harpe

Cape Peninsula University of Technology

South Africa

Isabella Venter

University of the Western Cape

South Africa

Gary Wills

University of Southampton

United Kingdom

(Reviewed paper)

DATA QUALITY: HOW THE FLOW OF DATA INFLUENCES DATA QUALITY IN A SMALL TO MEDIUM MEDICAL PRACTICE

ABSTRACT

Data is said to be of a required quality, if the data conforms to a defined specification and this specification correctly reflects its intended use. The importance of data and data quality within a Small to Medium Medical Practice (SMMP) cannot be ignored. Medical practitioners need accurate information in a form that is manageable and relevant to their context. The quality of data needs to be addressed in SMMPs as medical practitioners are often faced with making life threatening decisions.

The key elements identified in literature, that need to be considered before data quality can be understood are: the processes that generates and stores the data, data quality roles and responsibilities, the methods that exchange the data, complex nature of the data, flow of data and the different views of all the data stakeholders. The purpose of this paper is to determine the different data stakeholders of a SMMP and to establish the flow of the data between them. The flow of data is important in understanding the nature of the data.

The paper will depict the different data roles by applying a landscape model that will show the flow of data between data stakeholders in a SMMP. The three data roles of data producer, consumer and custodian are considered and to what extent the different data stakeholders perform these roles. This landscape model and the data roles can be used for further research to investigate the flow of data and how it impacts on the quality of data in a SMMP.

INTRODUCTION

Many organisations are drowning in data but they are unable to access information derived from this abundance of data. This becomes even more of an obstacle for organisations who quickly need to analyse large amounts of data from various sources accurately. The healthcare sector involve a myriad of stakeholders, including patients, health care providers, researchers, managed care organisations, third-party payers and medical doctors all of whom between them collect a large amount of patient data that is not integrated (Alshawi *et al*, 2003; Rosser & Kleiner, 1995). For example many of the patient data are stored in different formats and extracted from heterogeneous resources. The patient data could be extracted from paper files, electronic files, databases, spreadsheets and many other sources. This could lead to problems when real-time integrated information is required by the medical practitioner. Payton & Lucas (2001) are of the opinion that there is a need to have real-time access to information from many sources within the healthcare sector that could assist with decision-making of clinicians and support staff.

Medical practitioners accumulate an abundance of data from their patients during consultations. Although they have access to this data, it is under-utilised and not being used to its fullest potential (Long *et al*, 2004). When the patient moves to another practice then the patient data that was accumulated will be lost as most medical practitioners keep the file of a patient in the practice that collected it.

The quality of data in the healthcare sector is very important. Inaccurate data can lead to severe operational consequences that may influence life and death decisions. Data about the effectiveness of treatments, the accuracy of diagnoses and the practices of health care providers are crucial to improve health care delivery and to make better health care decisions (Leitheiser, 2001). The flow of data in the practice and the data stakeholders responsible for the data management has been identified as areas that could influence the quality of data.

FLOW OF DATA

The flow of data is important in understanding the nature of the data and focus on the sequence of activities from creation to disposition of data (Mathieu & Khalil, 1998; Strong *et al*, 1997). Data are the reflection of business objects and processes. Data evolve through a sequence of stages consisting of data collection, organization, presentation and application (Liu & Chi, 2002). The stages of the data evolution life can be further explained:

- *Data Collection* - data are captured through observing real world processes, measuring real world objects and perceiving real world stimulus.
- *Data Organization* - data are organized and stored in file-based data stores or databases.
- *Data Presentation* - data are processed, re-interpreted, summarized, formatted and presented in certain views.
- *Data Application* - data are utilized to achieve a certain application purpose, which in turn directs further data capturing.

According to Strong *et al* (1997) these activities have a direct impact on the quality of data. The flow of data and technologies contribute towards obtaining high quality performances and low defects. In a SMMP the creation and disposition of data need to be monitored to ensure that the data utilised is of a high quality. Data does not stay in one place according to Dravis (2004). Data moves in and out of a SMMP. In a SMMP patient data moves into the practice whenever the patient consults with the medical practitioner. The patient data can move out of the practice when it is sent to medical aids, specialists, hospitals and other third parties. These parties could return data back into the practice. Documenting the flow of data would indicate where the data is manipulated and when the usage of the data changes context. The flow of data is initiated by people and data in itself has no value except the fulfilment of purposes set forth by people. Due to the importance of the flow of data responsible individuals need to be assigned at the various stages of the data evolution life cycle.

DATA ROLES

The flow of data in any organization is initiated by people. Data itself has no value except to fulfill purposes set forth by people normally known as the data stakeholders. Rothenberg (1996) suggested that the quality of data should be established during the manufacturing of the data. Strong *et al* (1997) identified three roles within the data manufacturing cycle. The roles include data producers (people, groups or other sources who generate data and are associated with the data-production process), data custodians (people who provide and manage computing resources for storing and processing data and carry responsibility for the security of the data) and data consumers (people or groups who use the data, the people that utilize, aggregate and integrate the data).

Data producers generate data to meet a specification based on the need to represent some aspect of a defined reality. They conduct tests to validate the quality and accuracy of data. All data is produced with a purpose and the quality is based on the meeting of that purpose. The data producer will be responsible for the determination of data quality Rothenberg (1996). Strong *et al* (1997) added that the data custodian should take a broader conceptualization of data quality. Xu *et al* (2003) added a fourth data role, data managers, within the data manufacturing cycle. The data managers are responsible for managing data quality in the systems. Process owners should be made responsible for the quality of data that they produce in the organization (Methieu & Khalil, 1998). Different data roles might assign different priorities to data quality dimensions (De la Harpe & Roode, 2004). The literature highlighted that the flow of data and the various data roles should be considered when quality data is required within any organisation.

DATA QUALITY

It has been found that the users of data are unaware of the quality of data they use in their organizations (Parker, 2004). Data quality refers to how relevant, precise, useful, in context, understandable and timely data is (Firth, 1997, Barry & Parasuraman, 1997). Data is considered to be of high quality if it satisfies the requirements stated in a particular specification and the specification reflects the implied need of the user (De la Harpe & Roode, 2004). Data objects are said to be of a required quality, if the data conforms to a defined specification and this specification correctly reflects the intended use (Abate *et al*, 1998). According to Strong *et al* (1997) high quality data is data that is fit for use by the data consumers. The concept of quality is relative depending on the different perceptions and needs for the different data stakeholders.

Importance of data quality

Redman (1996) stated that poor data quality impacts a typical enterprise at various levels. At an operational level poor data quality leads to customer dissatisfaction, increased cost and lowered employee job satisfaction. Poor quality of data increases operational cost because time and other resources are spent detecting and fixing errors. The quality of data also impacts the data at a tactical level. At a tactical level poor data quality makes it more difficult to reengineer. At a strategic level data quality makes it more difficult to set and execute strategy. It also contributes to issues of data ownership and diverts management attention.

According to Dravis (2004) six factors or aspects of an organisation's operations should be considered. The six factors include context (type of data being cleansed and its purpose), storage (where the data resides), data flow (how data enters and moves through the organisation), work flow (interaction and use between work activities), stewardship (people responsible for managing the data) and continuous monitoring (processes for regularly validating the data).

The quality of data in the health sector is important and cannot be ignored (Parker, 2004). Inaccurate data can lead to severe operational consequences that may influence life and death decisions (Leitheiser, 2001). Grenson & D'Onofrio (2001) added that as in all organizations, the healthcare sector rely on data to manage and make better decisions. These activities depend on the quality of the data used to make them. Data quality could be described as a multi-dimensional concept with various characteristics depending on the view-point from the author. These various characteristics have also been described as data quality dimensions by many authors (Wang & Strong, 1996, Willshire & Meyeden, 1997, Eppler & Wittig, 2000).

Data quality dimensions

Extensive works have been done in the area of data quality frameworks to review areas where poor quality processes or inefficiencies reduce profitability of an organisation (Wang & Strong, 1996, Willshire & Meyden, 1997, Eppler & Wittig, 2000). A data quality framework should at least be used as a data quality assessment tool. Willshire & Meyden (1997) added that the data quality framework can go beyond basic assessment in an organisation. The framework can be used to model its data environment, identify relevant data quality attributes, analyse these attributes and provide guidance to improving data quality. Eppler & Wittig (2000) argued that a framework should also provide a scheme to proactive management data analysis.

Data quality defects are identified by comparing the information system with the represented part of the real world (Helfert *et al*, 2002) thus it is very important to have a clear picture of what the real world situation is. According to De la Harpe & Roode (2004) the social issues will escape attention and most probably thwart all attempts at bringing order to the data quality household. The flow of data and technologies also contribute towards obtaining high data quality performances and low defects (Cipriano, 1995).

The authors (Strong *et al*, 1997; Huang *et al*, 1999; Xu *et al*, 2002, Klein, 2002) identified the following data quality dimensions:

- *Intrinsic data quality* – Intrinsic data quality indicates that information has quality in its own right. It includes: accuracy, objectivity, believability, reputation, pragmatism, usefulness and usability.
- *Accessibility data quality* – Emphasizes that information system must be accessible but secure. Accessibility data quality includes: accessibility, access security and shared understanding of data by various social groups.
- *Contextual data quality* – Data that is provided in time and in appropriate amounts. Contextual data quality includes: relevancy, value-added, timeliness, completeness, amount of data and semantic.
- *Representational data quality* – Includes aspects related to the format of the information and its meaning. Representational data quality includes: interpretability, ease of understanding, concise representation, and consistent representation and syntactic.

According to Eppler & Wittig (2000) an information quality framework should achieve four goals. Firstly, the framework should provide a systematic and concise set of criteria to which the data can be evaluated. Secondly, provide a scheme to analyse and solve data quality problems. Thirdly, the framework should also provide the basis for data quality measurement and proactive management. Finally, it should provide a conceptual map that could be used to structure a variety of approaches, theories and data quality related phenomena.

RESEARCH METHODOLOGY

In order to come to a deeper understanding of the data quality problems of an organisation, it is also important to understand why an organization has quality problems in the first place and to what extent the environment contributes to these problems. To address this, it is necessary to study organizations in a real-life situation in an attempt to understand its data quality needs and problems (De la Harpe & Roode, 2004).

According to Yin (2002) case study research is used to study the contemporary phenomenon in its real-life context and it can be used where the research are at their early, formative stages. We therefore used qualitative case and analysis techniques as the data collection method. The medical practice was studied in its natural setting, the community it serves. Semi structured interviews were used to gather the data. Interviews were conducted with the key data stakeholders (people who have an interest in the data) of the practice which included data producers, data consumers and data custodians.

CASE STUDY ANALYSIS

Background

This SMMP is situated in a town called Vredenburg on the West Coast of the Western Cape. The community the practice serves is the middle to upper income area of Vredenburg. However, most of its patients come from the settlements a few kilometers from the practice. The SMMP consists of three doctors, a practice manager, an accounts official, a secretary and a creditor’s clerk. The practice has between 800 and

1000 patients visiting the practice per month. The practice uses both an electronic and paper patient record system.

Findings

Based on the definition of data roles by Strong *et al* (1997) and Xu *et al* (2002) we identified that all the data roles are present in the Vredenburg practice. The responsibility of the data roles are shared amongst the various data stakeholders in the practice. Some of the data stakeholders are responsible for more than one data role in a SMMP (see table 1).

Data Role	**Responsible**
Data Producer	Doctor, Patient, Secretary and Specialist
Data Custodian	Doctor and Secretary
Data Consumer	Doctor, Secretary, Account Official, Practice Manager and Creditors Clerk

Table 1: Vredenburg SMMP data roles

Only once different views and perceptions of all data stakeholders and the complex nature and flow of data are understood can data be utilized, which includes the concept of data quality (De la Harpe & Roode, 2004). A diagram indicating the flow of data in the SMMP was developed to identify the impact it has on data quality (see figure 1).

We identified three types of data that flow in a SMMP:

- Electronic data flow: The flow of data using technology (e.g. email, computers, databases and EPR systems).
- Paper data flow: The flow of data in the form of physical paper documents and files (e.g. test results, patient paper files and reports).
- Voice data flow: The flow of data through verbal communication (e.g. When a doctor verbally tell a patient his/her condition or diagnosis.).

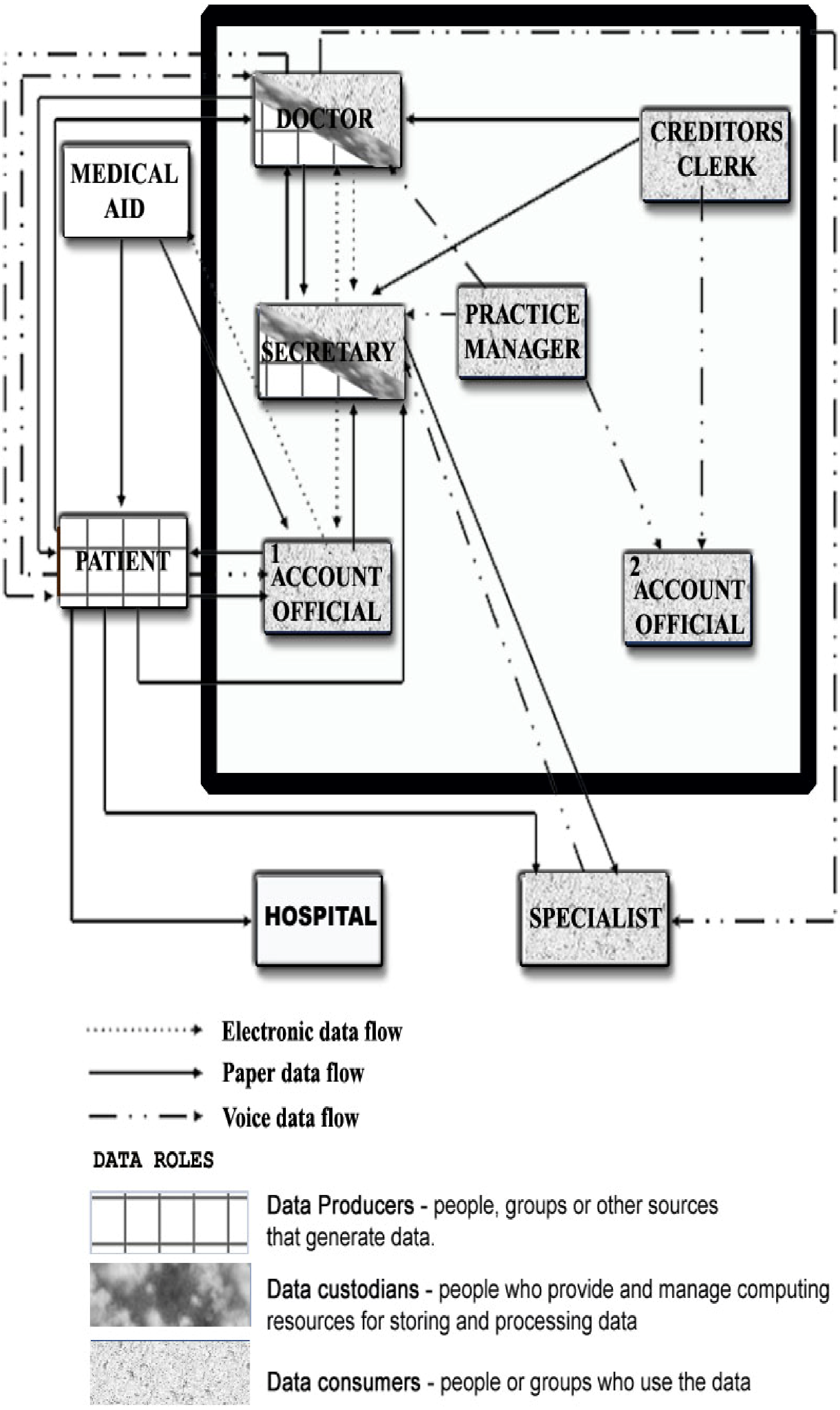
DOCTOR
CREDITORS CLERK
MEDICAL AID
SECRETARY
PRACTICE MANAGER
PATIENT
1 ACCOUNT OFFICIAL
2 ACCOUNT OFFICIAL
HOSPITAL
SPECIALIST
Electronic data flow
Paper data flow
Voice data flow
DATA ROLES
Data Producers - people, groups or other sources that generate data.
Data custodians - people who provide and manage computing resources for storing and processing data
Data consumers - people or groups who use the data

Figure 1: Vredenburg SMMP flow of data

The first point of data entry is when the patient visits the practice for the first time. The patient fills in a paper form and the secretary captures the data electronically. The secretary is responsible for validation of the data at first point of entry. There is a paper flow of data from the patient to the secretary. There are no user permissions set on any of the files, paper or electronic. The patient's data is only validated when the patient comes in for a consultation. If a patient does not come in often to consult with the doctor, his/her data is not used between consultations, the patients data may become out dated over time. In this practice electronic data is recaptured by the secretary after the patient filled in the data on the paper file and thus the electronic and paper data may differ from each other. The patient is responsible for filling in their own data for the patient file.

The patient data then moves to the doctor while the patient is in consultation. The doctor then captures diagnostic data on paper and electronically. The doctor verbally shares with the patient what the results are of the consultation. The patient also receives a note with the diagnosis and prescription. There is therefore a paper and verbal data transfer between the doctor and patient. The updated patient paper and electronic files are the transferred to the secretary. The paper files are stored in a cabinet with an indexed retrieval file system. The electronic file is on the network and there are no passwords to protect the patient data. Hence once an employee is logged into the network; they can access the patient files easily.

The secretary will electronically transfer the patient payment data to the accounts official. Data between a patient and account staff is via both electronic and paper transfers. The accounts employee receives patient data verbally from the creditor's clerk and doctor for confirmation of amounts. There is verbal transfer of data between practice manager and the secretary. If a patient qualifies for medical aid then the data is electronically transferred to the medical aid organisation by the accounts staff. The confirmation of medical aid is sent via paper to both the patients and medical practice.

Other flow of data in the practice occurs when a doctor refers a patient to either a specialist or hospital. There is a paper data flow between a patient and the specialist or hospital. The specialist may contact the doctor for more information and a voice data flow will occur between doctor and specialist. The patients are not allowed to view their data. The only access to their data is through the medical doctor verbally communicating it with them. There is therefore a verbal flow of data between the patient and the doctor.

Data quality issues identified

From the case study findings it is clear that the practice has a patient data *accessibility* problem. The reason for this is that the patient is at no point allowed to see their own files, so they can not view their own data. Another accessibility problem is that some of the reports and test results that specialist send back to the doctors are not captured electronically. The reports and test results are only filed in the form of a hardcopy document on the patient's paper file. Therefore certain data is only accessible in paper form and might not be readily available to the data consumer.

Due to the patient's data only being validated when the patient comes in for a consultation it leads to *timeliness* and *validation* data quality issues. For example a patient's address or contact details could have changed over time and the medical practice would not have means to contact this patient.

The data quality problem of *completeness* occurs because patients are responsible for the filing of their own data. The data that are captured by the secretary electronically is not validated for completeness. This could lead to data being omitted due to patients not knowing that it is important. *Security* data quality issues arise because there are no user permissions on the patient electronic data. This could lead to data being changed by unauthorised people and not being validated. Patient *confidentiality* could also be violated due to the lack of security in the practice. The data quality issues identified in this practice is accessibility, timeliness, completeness, security, validation and confidentiality. These issues have a direct impact on the quality of the data.

CONCLUSION

The literature clearly indicated that data without quality could cause various problems when the data is used for decision making. In healthcare the importance of data quality is crucial because the lives of patients are at risk. Within communities in the Western Cape SMMPs are the primary healthcare providers for the people. Medical practices therefore need to ensure that their data are of a required quality to support practitioners with decision making. The findings of the Vredenburg case study illustrated that, although the medical practice implemented a computerised patient record system, the practice still has data quality issues. The data quality issues identified in this paper proved to have a direct relationship with the flow of the data between the various data stakeholders in a SMMP. Further research in this area would be useful, to investigate how the data stakeholder landscape and flow of the data influence the quality of data in the practice.

References

Abate, M., Diegert, K. & Allen, H. (1998). A Hierarchical Approach to Improving Data Quality. *Data Quality* 4(1):365-369.

Alshawi, S., Missi, F. & Eldabi, T. (2003). Health care Information Management: the integration of patients' data. *Logistics Information, 16* (3/4): 286-295.

Barry, L.L. & Parasuraman, A. (1997). Listening to the customer – The concept of a service – Quality Information System. *Sloan Management Review* 38(3):65–76.

Cipriano, F. (1995). The Impact of Information systems on quality performance: An empirical study. *International Journal of Operations and Production Management* 15(6):69–83.

De la Harpe, R. & Roode, D. (2004). An Actor – Network Theory Perspective on Data Quality in Medical Practices. *Studies in Communication Sciences* 4(2):69-84.

Dravis, F. (2004). Data Quality Strategy: A step-by-step approach. *Proceedings of the ninth international conference on Information Quality (ICQ-04),*2004.

Eppler, M. & Wittig, D. (2000). A Review of Information Quality Frameworks from the Last Ten years. *Proceedings of the IQ 2000 – The Conference on Information Quality*. Boston, USA, 22-23October 2000.

Firth, C. (1997). *When do data quality problems occur?* Retrieved May 5, 2006, from http://www.sunflower-signet.com.sg/~firth/dql.htm.

Gendron, M. & D'Onofrio, M. (2001). Data Quality in the Healthcare Industry. *Data Quality*, 7(1).

Helfert, M., Zellner, G. & Sousa, C. (2002). Data quality problems and proactive data quality management in data warehouse systems. *Proceedings of the Business Information Technology World Conference*, 2002, 2 – 5 June.

Huang, K., Lee, Y. & Wang, R. (1999). *Quality Information and knowledge.* Upper saddle river, NJ: Prentice hall.

Klein, B.D. (2002). When do users detect information quality problems on the World Wide Web? *American Conference in Information Systems*, 2002.

Leitheiser, R. (2001). Data Quality in Health Care Data Warehouse Environments. *Proceedings of the 34th Hawaii International Conferences on System Sciences. Hawaii*, 2001.

Long, J. Seko, C., Robertson, C & Morrison, L. (2004). Where to start? A preliminary data quality checklist for emergency medical services data. *Proceedings of the ninth international conference on Information Quality (ICQ-04).*

Liu, L. and Chi, L. (2002). Evolutional Data Quality: A Theory-specific view. *Proceedings of the Seventh International conference on Information Quality (ICQ-02).*

Mathieu, R. & Khalil, O. (1998). Data quality in database systems course. *Data Quality Journal* 4(1).

Parker, M. (2004). A generic business intelligence data model to analyse data within a small to medium medical practice (SMMP). Conference Paper. *South African Institute of Computer Scientists and Information Technologists (SAICSIT)*: Stellenbosch, South Africa, October 2004.

Payton, F. & Lucas, H. (2001). Health Care B2C Electronic Commerce: What do Patients Consumers want? *Proceedings of the 34th Hawaii International Conferences on System Sciences*. Hawaii.

Rosser, L & Kleiner, B. (1995). Using Managements Information Systems to enhance health care quality assurance. *Journal of Management in Medicine, 9* (1): 27-36.

Rothenberg, J. (1996). Metadata to Support Data Quality and Longevity. IEEE Software Retrieved May 21, 2006 from http://www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html

Strong, D., Lee, Y. & Wang, R. (1997). Data Quality in context. *Communications of the ACM* , 40(5): 103-10.

Wang, R. & Strong, D. (1996). Beyond accuracy: What Data Quality means to data consumers. *Journal of Management Information Systems*, 12(4): 5-34.

Willshire, M. & Meyden, D. (1997). A process for improving data quality. *Data Quality Journal*, 3(1).

Xu, H., Nord, J., Brown, N. & Nord, G. (2002). Data quality issues in implementing an ERP. *IEEE transactions on knowledge and data engineering*, 1021(1):47-58.

Yin, R. (2002). *Case study research, Design and methods*. 3rd edition Newbury Park. CA: Sage Publications.

Acknowledgements

The authors would like to acknowledge Cape Peninsula University of Technology, National Research Foundation, University of the Western Cape, University of Southampton and INDEHELA for their support and contributions.