SNAP-ON DATA QUALITY ENHANCEMENT AND VERIFICATION TOOL (DEVA) FOR ASSET MANAGEMENT

Jing Gao
University of South Australia, Australia
jing.gao@unisa.edu.au

Andy Koronios
University of South Australia, Australia
andy.koronios@unisa.edu.au

Abstract: During the data collection process, human error is a large reason asset management organisations suffer from poor data quality (specifically, as not all data entries can be replaced by the automatic acquisition method). Thus, to reduce human error as one cause of data quality problems, software assistance is considered valuable. This paper provides an innovative client-server based software solution. By using the operating system interception method, this solution can transfer the data quality (business) rules generated from the data sets and apply them to the existing information system at the client side real-time. This snap-on approach has great potential to be applied in domains other than asset management, resulting in great enhancement of data quality.

Key words: Data Quality, Information Quality, Snap-on data quality software

INTRODUCTION

Gaining control of assets is clearly challenging. Asset data and information is a key enabler in gaining control of assets. Asset data and information is created by many departments, in many forms, during all stages of a typical asset's lifecycle and thus is complex to manage. These data and information are likewise used by a diverse set of people and systems, each with their own specific needs and requirements. Consistent and reliable data and information is essential for both operations and maintenance activities.

Similar to other business areas, data and information that is comprehensive, accurate and immediately accessible enables people to make decisions faster and more accurately, leading to higher availability and lower maintenance costs. Gaps in asset information, out-of-date or wrong information, or the inability to rapidly access necessary information, wastes time and money and reduces return-on-assets. In many instances, the lack of information needed for good decision-making may result in no action being taken, with disastrous consequences (Frame 2003).

Because of the importance of asset information in gaining control of assets, the quality of asset information becomes critical in managing the asset's availability and reliability. Quality asset data provides the foundation for effective asset management and optimised asset management decisions. However, a previous study on data quality in asset management organisations (Lin et al, 2007) suggests that the majority of asset management organisations face issues in data quality at all organisational levels. This study further points out that, when there is a doubt about the quality of data obtained, most data consumers cross-check (80%), ask the field people (71%), or make assumptions based on their own experience (67%). These verification activities are usually expensive in time and financial resources. For example, many data consumers would capture the data personally, check onsite themselves or conduct field checks themselves. It must also be noted that this also implies data consumers have a lack of trust in the data collectors in the field.

With respect to the various data quality issues in asset management organisations (and possibly in other organisations), this paper tries to propose an innovative approach to enhance data quality from the entry point (by data collectors) in accordance with the data consumer's business requirements.

DATA QUALITY

Managers intuitively differentiate information from data, and describe information as data that has been processed. However, data and information are often used synonymously in practice, particularly when addressing quality issues. Therefore, this paper (and the survey) uses “data” interchangeably with “information”, as well as using “data quality” (DQ) interchangeably with “information quality” (IQ).

There are a number of theoretical frameworks for understanding data quality. These DQ frameworks (Wand 1996, Wang 1996, Wang 1998, Price and Shanks 2004, Eppler 2001, Giannoccaro 1999, etc) have been proposed to organise and structure important issues in information quality, albeit from different points of view. This paper follows Wang & Strong’s (1996) data quality definition and regards the quality of data as being multi-dimensional, including as it does accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness, and usefulness. In addition, it is also considered that the quality of data is dependent on how the data will be used (e.g. Ballou and Pazer 1995, Neely 2001 and Strong 1997). This fitness for use can be defined as the intersection of the quality dimension(s) being considered, the proposed use of the data (purpose), and the data fields which are identified for use in order to fulfil the purpose (Neely and Pardo, 2002).

The quality of the data that managers use is critical. Without quality data, organisations are “running blind”; making a decision becomes a gamble (ARC 2004). The lack of quality data often leads to decisions being made more on the basis of personal judgment rather than being data driven (Koronios, Lin et al 2007). Poor data quality can diminish the value of otherwise successful systems of many kinds, including data warehouses and enterprise resource planning (ERP). It is also likely to create risks for an organisation in ensuring regulatory compliance such as with the Sarbanes-Oxley Act in U.S. and Basel II in Europe. Poor quality data, if not identified and corrected, can have disastrous impacts on the health of the organisation (Wang & Strong 1996). These impacts range from operational inconvenience to ill-informed decision-making, to disruption of business operations, and possibly even to organisational extinction.

DATA ACQUISITION METHODS IN ASSET MANAGEMENT ORGANISATIONS

Asset information is created throughout all stages of a typical asset’s lifecycle; thus managing the flow and the quality of information is critical to managing the asset’s availability and reliability (ARC 2004; Paiva et al 2002). In practice, data are captured both manually through computers, PDAs, paper forms etc, and/or automatically through sensors and transducers. It is acknowledged by many studies that data quality issues are found in both automated and manual data acquisition methods, as discussed below.

Automated Data Acquisition

Automated data acquisition methods are generally considered to provide higher quality data than those requiring human intervention (Smith 2002). While there may be improvement in some dimensions, there are other considerations: sensors can be faulty or out of calibration; and their data tend to be unreliable (Jeffrey 2006). Reliance on sensor data can give a false sense of confidence in the captured data. This is often mitigated in critical systems by building in redundancy - adding a second sensor to verify the accuracy of the first.

Asset management data can come from constant or transient data streams. These can be continuous, “multiple, rapid, time-varying, possibly unpredictable and unbounded streams” (Babcock 2002). Recording stream data in a database for subsequent analysis requires values to be sampled at some pre-determined rate and the sheer volume of data may become difficult to analyse and store.

Electronic sensors or transducers are used for condition monitoring. As captured signals are generally very weak, a charge amplifier is connected to the sensor or transducer to minimize noise interference and prevent signal loss. The amplified signal is then sent via coaxial cable to a filter for noise removal and routed to a signal conditioner. The signal is used to visually indicate the physical quantity being measured but needs conversion to digital form before it can be stored or accessed by other information systems. The precision of an analogue to digital (A/D) converter must match the precision to which the data needs to be captured. Choosing a lesser number of bits for the A/D converter to represent the analogue signal may result in low precision of recorded data that may not be suitable for some uses of the captured data. To maximise the benefits of automated data capture, careful consideration must be given to the data sampling rates; the required precision; the criticality of the component; and the potential consequences if its related data is not correct, complete, and/or timely.

Manual Data Acquisition

Asset management organisations employ a wide array of specialists to install, assess, service, maintain, and upgrade their assets. It is critical to record their activities, judgements, and interpretations in order to manage assets efficiently (ARC 2004). Reliance on human operators to provide this data introduces many potential sources of degradation to its quality, including threats to its correctness, completeness, and timeliness.

Human error is a large reason asset management organisations suffer from poor data quality and not all data entries can be replaced by the automatic acquisition method. Thus, to reduce human error as one cause of data quality problems, software assistance is considered valuable.

CURRENT DATA QUALITY TOOLS

Pohlmann (2004) and Knightsbridge (2006) predicted that the information quality market for software and professional services will reach US$1 billion by 2008, indicating that organisations are taking these potential DQ problems quite seriously and are investing in solutions. Data quality software vendors are offering a wide range of data quality functionality, like data profiling, data parsing/ correction, data matching/de-duplication, enrichment, integration, and data monitoring (Howard, 2004). They are either offering various data quality components as a separate product, with some degree of integration between them, or a suite of functions covering the full spectrum of capabilities. So it becomes convenient for organizations to deploy a single-vendor solution for organisation-wide data quality requirement. Table 1 shows a list of data quality tools that are currently available.

Table 1: Example of data quality tools

Data Quality Tool	**Vendor/ Developer**
AJAX	INRIA, France
Arktos	National Technical University, Athens
Athanor	Similarity Software
ChoiceMaker	ChoiceMaker Technologies
DataLever Enterprise Suite	DataLever Corporation
DataMapper	Exeros
DataSight	Group 1 Software – Pitney Bowes
dfPower Studio	DataFlux SAS
Dn:Clean	Datanomic
FactoryTalk	Rockwell Automation

i/Lytics	Innovative Systems Inc
Information Quality Suite	First Logic – Business Objects
Intelliclean	National University of Singapore
OptimizeIT Data Manager	ABB
Porter’s Wheel	University of California, Berkely
PowerCenter	Informatica
Telcordia	Telcordia Technologies
Trillium Software System	Harte-Hanks Trillium Software
WebSphere QualityStage Enterprise Edition	IBM
WinPure	WinPure

Having reviewed the features of these software offerings, it is found that nearly all these tools are solutions for discovering and generating data quality rules for treating data quality problems in the existing data repository (lakes), but are not effective in tackling the source problems (rivers).

RESEARCH DESIGN

This research aims to develop a software-assisted information quality assurance method for engineering asset management organisations, which will allow engineering asset data to be analysed, profiled and cleansed, as well as enriched, to ensure multiple and meaningful uses of such data. In particular, a specific focus has been placed on how to better design the information systems to allow for quality data entries. This research consists of two stages:

- Stage One: what are the issues associated with the current information systems in asset management organisations that prohibit quality data entries?
- Stage Two: What is an adequate software solution to address the issues determined during stage one and potentially result in high data quality?

During Stage One, the researchers have conducted a number of interview-based case studies with multiple asset management organisations. Findings are summarised in the next section. During Stage Two, a prototype software solution was designed to address the identified issues.

FINDINGS

Human Errors

The simplest and most obvious cause of poor data quality is human error. The characteristics of the data and their capture and acquisition procedures, processes, and the environment common to asset management, are especially conducive to the following:

- The asset operators are typically highly trained maintainers and engineers, but not data-entry specialists. Such operators can make typographical or transcription errors and enter incorrect values affecting the correctness dimension of data quality. They are not likely to be familiar with the intricacies of complex asset management data-entry systems and may miss important fields, negatively affecting their completeness.

- The environment where assets are installed or located may be harsh. Reading and recording of the data in such situations may be difficult. The operator may misread the value at the source and enter what is believed to be the correct value into the system; however, the entered data may be incorrect.

- The data entry location may similarly not be conducive to accurate data entry. It may be busy or noisy, leading to correctness problems. The data source is sometimes separate from the data entry location, forcing personnel to rely on their memory, or to make notes for later transcription. This can negatively affect many dimensions of data quality, not the least of which is timeliness. There is an inherent delay between the time the data is read and the time it is recorded in the system. Depending on many factors, including distance between sites, personnel workload, and availability of data-entry workstations to name just a few, notes may become lost or indecipherable and memory may not be very reliable, resulting in incorrect values being entered (correctness) or missed out altogether (completeness).

- Personnel are frequently transferred from one site or department to another, potentially reducing their sense of ownership of, and pride in, the quality of their work, as well as their familiarity with the system used in each location. Resultant errors and problems are often left to become the responsibility of their successor, further compounding the problems. Many observations made by human agents are in the form of unstructured data. Such data are very difficult to validate at the time of capture or acquisition, leading to a reduced sense of confidence in its accuracy when it is later processed.

Business rules constraints

During the interviews, data quality related business rules were addressed several times. In many cases, the interviewees suggested that by developing business rules and applying them within the information systems (especially at the entry point), the data quality can be enhanced. However, in practice, the following issues were found to prevent business rules being effective:

- Unable to discover business rules from the existing information system where documentation may have been lost or the source codes are not available
- Using data mining techniques to discover business rules in asset management database systems is possible. However, without expert human verification, the rules can be meaningless and duplicated.
- Applying business rules into the existing systems can be expensive and time consuming. This often requires systems re-engineering. In big organisations, the new release of existing systems may take years. By the time the new system is in place, the business rules may be out-of-date.

PROPOSED SOLUTION

The proposed data quality tool is a client-server based solution. The client side focuses mainly on detecting problems with input data at the time of the data capturing process. The client side (a software agent) is designed to intercept the data entered (user operating system interface interception and web-based and windows-based application process - id hooking) and test it in real time for errors and inconsistencies through comparison to business rules. When an error has been detected, the "submit" button will be disabled unless corrections are made (or the system may simply give a warning message). Otherwise, the client agent will sit in the background quietly. Further the client side will download the latest business rule definition files from the server-side and apply them to the desktop systems in real-time.

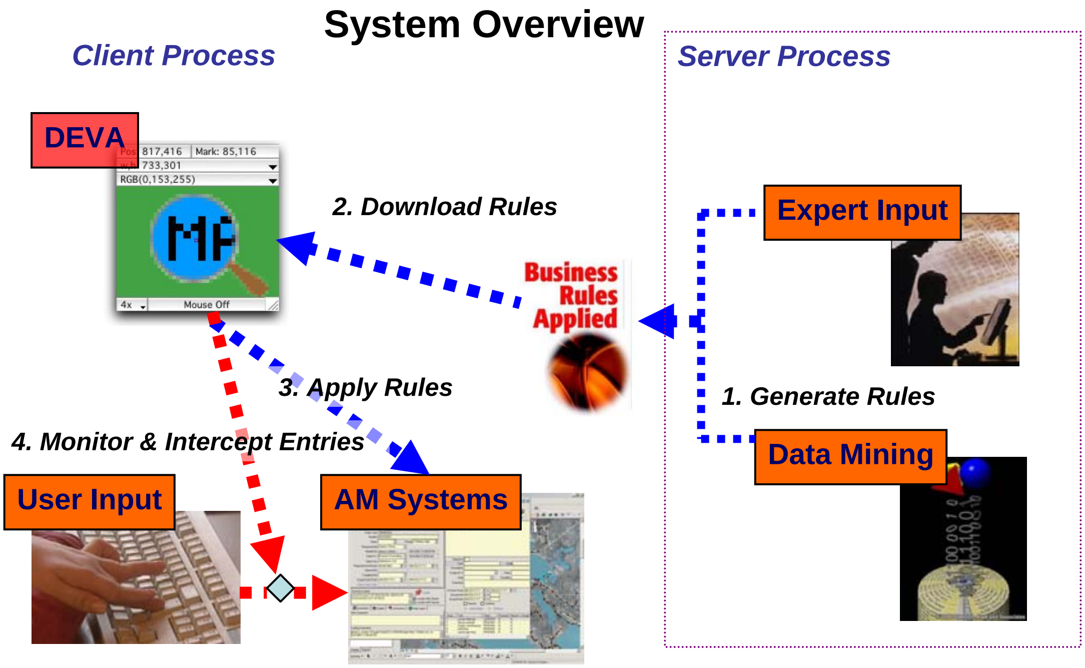
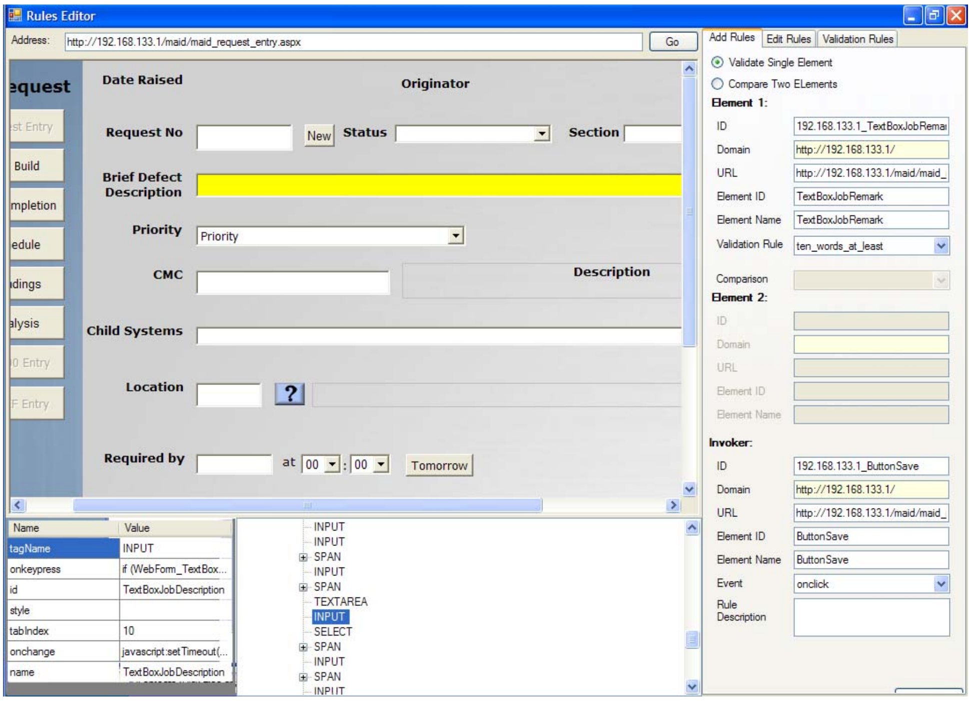


Figure 1: DEVA system overview

In data analysis, profiling is a resource-intensive exercise. Thus, this exercise is completed at the server-side where a data mining application can go through the existing data sets to generate potential business rules. Once these rules are verified by the human experts, the rule definitions file (XML-based) will be created and delivered to the client side (the best analogy may be found in the process currently adopted by anti-virus software). A more complicated rule will be created as DLL binary files.

The term “rule” needs to be clarified. There are two kinds of rules in the proposed system: interception rules and validation rules. A validation rule (business rule) consists of description, format and warning elements. Description is a short text to describe the purpose or form of the validation. Format is a regular expression for correct input value. Any value does not match this regular expression is considered to be invalid. Warning is the text to be displayed when a validation is failed. The regular expression can be derived from the data mining exercise. With human expert's verification, the regular expression can be converted to a complete validation rule. On the other hand, an interception rule describes what element should be evaluated against what validation rule at what event (e.g. menu selection, textbox entry, etc). So an interception rule always contains at least one validation rule.

Writing all interception rules by hand can be time-consuming. So a rule editor is included to assist defining new rules. For a web application, a rule editor is provided as shown below.



Address: http://192.168.133.1/maid/maid_request_entry.aspx

Priority: Priority

Required by at 00 : 00

Name	Value
tagName	INPUT
onkeypress	if (WebForm_TextBox...
id	TextBoxJobDescription
style	
tabIndex	10
onchange	javascript:setTimeout(...
name	TextBoxJobDescription

● Validate Single Element
○ Compare Two ELements

Element 1:

ID	192.168.133.1_TextBoxJobRemar
Domain	http://192.168.133.1/
URL	http://192.168.133.1/maid/maid_
Element ID	TextBoxJobRemark
Element Name	TextBoxJobRemark
Validation Rule	ten_words_at_least
Comparison	

Element 2:

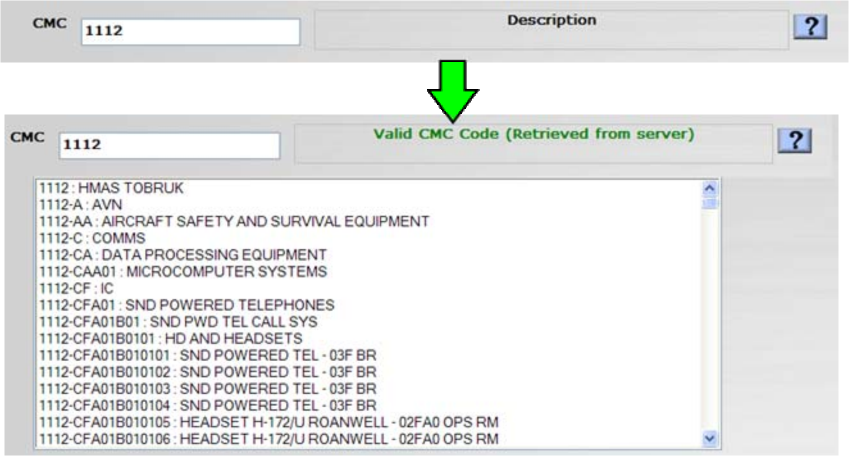
ID	
Domain	
URL	
Element ID	
Element Name	

Invoker:

ID	192.168.133.1_ButtonSave
Domain	http://192.168.133.1/
URL	http://192.168.133.1/maid/maid_
Element ID	ButtonSave
Element Name	ButtonSave
Event	onclick
Rule Description	

Figure 2: DEVA Web Application Rule Editor

The DEVA system is designed to deal with multi-dimensional data quality problems. For example, when a user enters the maintenance hours for a particular job, the server side will provide an indicative range for this particular job type based on the average figure calculated through the data mining exercises. Additionally, the software can also be used to provide assistance for data collectors. The example below shows that the original system requires a manual entry of CMC code (asset id); once the DEVA client is activated, the text box has been converted to the pull down menu, which gives the data collector a good idea of what the correct format is.



CMC 1112

CMC 1112 Valid CMC Code (Retrieved from server)

1112 : HMAS TOBRUK
1112-A : AVN
1112-AA : AIRCRAFT SAFETY AND SURVIVAL EQUIPMENT
1112-C : COMMS
1112-CA : DATA PROCESSING EQUIPMENT
1112-CAA01 : MICROCOMPUTER SYSTEMS
1112-CF : IC
1112-CFA01 : SND POWERED TELEPHONES
1112-CFA01B01 : SND PWD TEL CALL SYS
1112-CFA01B0101 : HD AND HEADSETS
1112-CFA01B010101 : SND POWERED TEL - 03F BR
1112-CFA01B010102 : SND POWERED TEL - 03F BR
1112-CFA01B010103 : SND POWERED TEL - 03F BR
1112-CFA01B010104 : SND POWERED TEL - 03F BR
1112-CFA01B010105 : HEADSET H-172/U ROANWELL - 02FA0 OPS RM
1112-CFA01B010106 : HEADSET H-172/U ROANWELL - 02FA0 OPS RM

Figure 3: DEVA in action

CONCLUSION

Data quality has been an acknowledged issue for a long time. There is strong evidence that most organisations have far more data than they can possibly use; yet, at the same time, they do not have the data they really need. Many say “we are drowning in data and are starved of information”. Despite this apparent explosion in the generation of data it appears that, at the management level, executives are not confident that they have enough correct, reliable, consistent and timely data upon which to make decisions.

Maintaining the quality of data is often acknowledged as problematic, but is also seen as critical to effective decision-making. In practice, the data quality issues are often attributed to the human errors occurring during manual data acquisition, as has been identified during the interviews. Although many studies show that building data quality rules into the existing information systems may improve data quality, in practice (due to the long information system upgrade/change cycle, or out-sourcing limitations), these initiatives are not implemented.

Based on these findings, this research has provided a prototype client-server based software solution. By using the operation system interception method, this solution can transfer the data quality (business) rules generated from the data sets and apply them to the existing information system at the client side real-time. This snap-on approach has great potential for applications in domains other than asset management, greatly enhancing data quality.

REFERENCES

[1] ARC 2004, Asset Information Management – A CALM Prerequisite, ARC Advisory Group, Boston, USA

[2] Ballou, D.P., & Pazer, H.L. (1995). Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. Information Systems Research. 6(1), 51-72.

[3] Eppler, M.J. (2001). The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks. Studies in Communication Sciences. 1, 167-182.

[4] Frame, J. Davidson, 2003. Managing Risk in Organizations: A Guide for Managers. Jossey-Bass, San Francisco, CA.

[5] Giannoccaro, A., Shanks, G., & Darke, P. (1999). Stakeholder perceptions of data quality in a data warehouse environment. Australian Computer Journal. 31(4), 110-117.

[6] Howard, P., (2004), Data Quality Products: an evaluation and comparison, Bloor Research Report, Bloor Research, Milton Keynes, United Kingdom.

[7] Jeffery, S.R., Alonso, G., Franklin, M.J., Hong, W. and Widom, J., 2006, 'Declarative support for sensor data cleaning', paper presented at the 4th Pervasive Conference.

[8] Knightsbridge 2006, Top 10 Trends in Business Intelligence for 2006, Knightsbridge Solutions LLC, Chicago, IL.

[9] Lin, S., Koronios, A., & Gao, J., 2007, “A Data Quality Survey for Asset Management in Australian Engineering Organisations”, The 5th International Supply Chain Management and Information Systems Conference (SCMIS 2007), 9-12 December 2007, Melbourne, Australia.

[10] Neely, P., & Pardo, T., “Teaching Data Quality Concepts Through Case Studies”, Center for Technology in Government, Albany, 2002.

[11] Neely, P., “A Framework for the Analysis of Source Data Revised”, Proceedings of AMCIS 2001.

[12] Pohlmann, T 2004, 2005 Enterprise IT Outlook: Business Technographics North America, Forrester.

[13] Price, R.J. and Shanks, G. (2004). A Semiotic Information Quality Framework. Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference 2004. 1-3 July 2004, Prato, Italy. 658-672.

[14] Smith, A.D., Offodile, F. (2002). Information management of automatic data capture: an overview of technical developments. Information Management and Computer Security. 10(2), 109-118.

[15] Strong, D.M., “IT Prosess designs for Improving Information Quality and reducing Exception Handling: A Simulation Experiment”, Information and Management, (31), 1997, pp. 251-263.

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

[17] Wang, R., “A product perspective on data quality management”, Communications of the ACM, 41(2), 1998, pp. 58-65.

[18] Wang, R.Y., and Strong, D.M., “Beyond Accuracy: What Data Quality Means to Data Consumers”, Journal of Management Information Systems, 12(4), 1996, pp. 5-33.

ACKNOWLEDGEMENT:

The work of this paper is supported by CIEAM CRC, Australia.