Examining Data Cleansing Software Tools for Engineering Asset Management

Vivek Chanana, University of South Australia, Adelaide, SA 5001, Australia; E-mail: Vivek.Chanana@unisa.edu.au
Andy Koronios, University of South Australia, Adelaide, SA 5001, Australia; E-mail: Andy.Koronios@unisa.edu.au

ABSTRACT

To be more cost effective and efficient, organizations are relying on improved operations and maintenance strategies for their engineering assets. Even after having done huge investment in operations and maintenance systems, they are not able to reap proportional benefits. It is mainly attributed to the quality of data present in these systems and lack of integration between diverse systems used in the organizations. The paper presents uniqueness of data encountered in engineering asset management (EAM) setup and the quality problems in such data. It reviews the suitability of commercially available data cleansing software tools for EAM environment, provides a comparative evaluation of features of tools and points to the areas that need improvement to effectively handle EAM data.

Keywords: Engineering asset management, data cleansing, data quality, data cleansing tools.

1. INTRODUCTION

Striving for success in today's global economy has put manufacturing and utilities organizations under intensive pressure. They are trying to achieve this through increasing production efficiency, reducing inventory carrying costs through just-in-time strategies, while ensuring that their products meet customer satisfaction. Unfortunately, organizations have not done equally well on the asset management front. To achieve maximum possible up-time for their assets, organizations are now trying to turn towards asset management as an optimization strategy to improve their process efficiency, reduce maintenance cost, and improve their returns over time on assets (Eerens, 2003). The objective of effective engineering asset management (EAM) is to reduce total cost of ownership of assets, ensure the optimum level of reliable and uninterrupted delivery of quality service, minimize the need for new assets, and continuously align assets with organizational needs (IPWEA, 2002).

Engineering asset management involves acquiring, maintaining, and disposing of assets and is a serious business. Staggering investments are made in engineering assets. The cost of maintenance and replacement of these assets represents a major share of their operating costs. According to a study conducted by ARC Advisory Group (Snitkin, 2003), nearly 40 percent of manufacturing revenues are budgeted for maintenance. The potential for savings from proper management of assets is immense. Obviously, a little improvement can result into significant savings through reduction in maintenance cost. Enhancing quality of data stored in various information systems supporting EAM through data cleansing can reduce the operational and maintenance cost in organizations and contribute significantly to their bottom line.

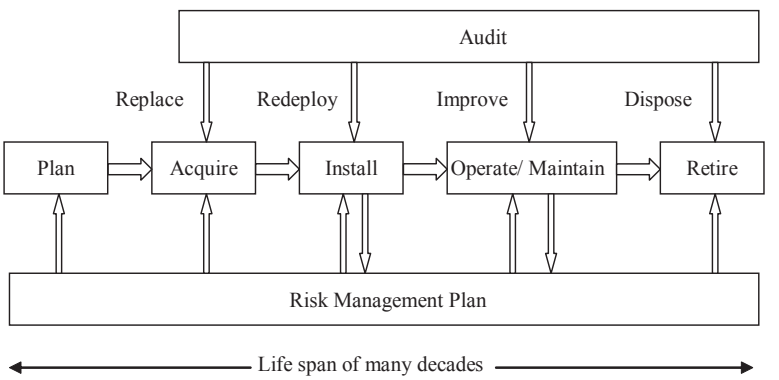
In this paper, the authors discuss the results of a study done to evaluate commercially available data cleansing software tools for EAM environment. The paper is organized as follows. Section 2 discusses how data encountered in EAM is different from data found in business applications. Section 3 defines the concept of data quality and highlights the quality problems in such data. Section 4 then provides a solution to improve data quality through data cleansing. Section 5 provides a list of commercially available data cleansing tools that were investigated and also provides a comparative evaluation of features of these tools and points to the areas that need improvement to effectively handle EAM data.

2. EAM ENVIRONMENT

Engineering asset management deals with the effective management of broad range of physical assets like machinery, production equipment, and fleets belonging to engineering organizations. Engineering assets exhibit unique characteristics in many ways. They are generally complex, expensive and operate for continuous periods of time. They need care so that they can provide service without failure for an extended useful lifetime. They need to be managed through best practice methodologies and business processes for getting maximum benefit from them (Blanchard, 2006). The process of asset management is sophisticated and involves the whole asset lifecycle that generally spans over many decades as shown in Figure 1.

Maintenance strategies for engineering assets, once *run-to-failure* now are becoming *condition-based*. Condition-based maintenance used in strategies like Reliability

Figure 1. Engineering asset lifecycle stages



Centered Maintenance (RCM) constantly monitor the health of assets and collect lots of data pertaining to asset conditions. In addition to this, data related to asset operations are also collected. The analysis of collected data provides knowledge about the current and future condition of assets and helps to schedule and plan future maintenance activities.

Data typical of asset management environment can be both structured and unstructured. The structured data could be related to physical characteristics of assets like its type, size, ratings, specifications, bill of materials (BOM) of parts and spares etc. The unstructured data could in the form of notes, inspection reports, sketches, work instructions, safety data sheets etc. Data can be captured both automatically through sensors and field devices, and manually through human operators and technicians in a variety of formats. The automatically captured data can be periodic or continuous process-centric streaming real-time data. Data could also be sourced from a variety of databases and disparate operation and maintenance systems.

Data types can typically include: inventory data, condition data, performance data, criticality data, lifecycle data, financial data, risk data, reliability data, technical data, physical data, GPS data, etc. EAM environment contains many information systems used at various life cycle stages and these systems store data in their propriety formats. Some of the systems used in typical EAM environment are:

- CAD: Computer Aided Design
- CMMS: Computerized Maintenance Management System
- DMS: Document Management System
- ERP: Enterprise Resource Planning
- GIS: Geographic Information System
- LIMS: Laboratory Information Management System
- PAM: Plant Asset Management
- PDM: Product Data Management
- PLM: Product Life Cycle Management
- SCADA: Supervisory Control and Data Acquisition
- SRM: Supplier Relationship Management
- WMS: Warehouse Management System

3. DATA QUALITY ISSUES

Data quality has become an important topic of investigation in research and industry however, so far there is no single agreed definition of data quality (Maletic and Marcus, 2000). We define high-quality data as data that are *fit for use* by data consumers—a widely adopted criteria (Strong et al, 1997). Data quality is regarded as multidimensional concept in the literature and there is no general agreement on a firm set of data quality dimensions (Wand and Wang, 1996; Wang et al, 1995). Frequently mentioned dimensions are:

(i) accuracy (degree of correctness and precision with which real world data of interest to an application domain are represented in a system),
(ii) completeness (degree to which all data relevant to an application domain have been recorded in a system),
(iii) consistency (degree to which the data managed in an information system satisfy specified constraints and business rules), and
(iv) timeliness (degree to which the recorded data are up-to-date).

Data used in EAM organizations can have a wide range of errors, inaccuracies and inconsistencies, such as wrong data, missing data, inconsistent use of abbreviations, misspellings during data entry, outdated or invalid data etc. in single or multiple data sources. Also EAM organizations generally have a significant portion of their data stored in legacy systems that constitutes lots of unstructured data. These organizations face many problems in mapping field names while migrating from legacy systems to enterprise asset management applications.

As discussed in the last section, EAM organizations use plethora of information systems that deal with asset data. Data in these information systems have great diversity in formats and semantics (Friedman, 2006). These organizations cannot deliver attractive return on investment unless they are underpinned by clean and consistent data. As systems grow in complexity and the volume of data increases, the level of data quality becomes more critical to success. The effectiveness of decision-making is limited to the quality of the data. A number of organizations worldwide have suffered large financial losses due to inaccurate, incomplete or wrong data in their data repositories (English, 1999).

4. DATA CLEANSING

Data cleansing deals with detecting and removing errors and inconsistencies from data in order to improve their quality (Muller and Freitag, 2003; Rahm and Do, 2000). Data cleansing also called data cleaning or scrubbing has caught on in a big way as a crucial first step in organizations dealing with applications like data warehousing, data mining and knowledge discovery where data forms their core asset.

As mentioned in section 2, asset management is characterized by a range of diverse data sources most often from multiple vendors and at times with some proprietary data formats. Data cleaning becomes especially important in such situations when multiple data sources need to be integrated. On the lower end of the problem, it could be a case of simple redundant data that can be handled easily. On the other end, semantically same data having different representations in heterogeneous sources will result into redundant data. In order to have a single view and access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information becomes very necessary. This problem has been referred to in literature as merge/purge problem (Hernandez and Stolfo, 1998).

While most practitioners of data quality are aware of the problems with their data quality issues, only recently there has been an emphasis on the systematic detection and removal of data quality problems (Dasu and Johnson, 2003). A data cleansing approach detects and then removes all major errors and inconsistencies both in individual data sources and when multiple data sources are integrated. Therefore, data cleansing should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata.

5. DATA CLEANSING TOOLS

5.1 Commercially Available Data Cleansing Tools

In recent years there has been a significant growth of data quality tools in the market. This is the result of organizations realizing the importance of good quality data and the harm poor quality data can do to their effectiveness. The organizations are considering data quality as a strategic issue and are actively seeking solutions for its improvement. They are relying on technology for improving data quality which has led to a strong interest in data quality tool market. Although data quality tools market is still modest in size compared to many other software markets, there is a likelihood of it growing in future. This is evident by the tumultuous movement in the market and the interest it has evoked in small and large vendors. There have been mergers and high profile acquisitions by large vendors like Business Objects, IBM and Pitney Bowes.

The authors investigated the suitability of some commercially available data cleansing software tools in the market for the cleansing of asset management data. Our analysis is based on our appraisal of the available literature found on vendor's homepages, their product specifications, and independent industry survey reports from consulting firms. The limitation of this research at this stage is that the comparison of the cleansing tool is not based on any testing of installed software with real data. The reason for not doing this is that the cleansing software tools are very expensive and hence it was not practically possible and financially feasible to investigate all software installations. In future research activities we plan to interview the vendors so that we can get information that is not freely available. Later on, with the identification of industry partners of our University who have used these software tools, we plan to have even more in-depth viewpoint and feedback about the respective advantages and features of these tools. Although every attempt was made to include major vendors in this study, it is possible to have missed some major vendor in this area and/ or not adequately covered some specification of their tools that they might support.

Following is the list (in alphabetical order) of commercially available tools that we reviewed:

- Athanor - Similarity Software (IMC, 2006)
- DataSight - Group 1 Software - Pitney Bowes (G1S, 2006)
- dfPower Studio - DataFlux SAS (DFC, 2006)
- i/Lytics - Innovative Systems Inc (ISI, 2006)
- Information Quality Suite - First Logic - Business Objects (FLI, 2006)
- PowerCenter - Informatica (IMC, 2006)
- Trillium Software System - Harte-Hanks Trillium Software (HTS, 2006)
- WebSphere QualityStage - Enterprise Edition IBM (IBM, 2006)

Table 1. Comparison of data cleansing software tools

Features	Athanor 3.0	DataSight	dfPowerStudio	i/Lytics	IQ8	PowerCentre	Trillium Software System	WebSphere QualityStage
Vendor	Similarity Systems (Informatica)	Group1 Software (Pitney Bowes)	DataFlux SAS	Innovative Systems	First Logic (Business Objects)	Informatica	Harte-Hanks Trillium Software	IBM
Sub systems	Athanor Designer Athanor Server Athanor Runtime Athanor RealtimeSDK	CODE 1 Plus Suite Universal coder Merge/Purge Plus	dfPower Profile dfPower Quality dfPower Customise	i/Lytics Data Profiler i/Lytics Data Quality i/Lytics GLOBAL i/Lytics SECURE			TS Discovery TS Quality	DataStage QualityStage ProfileStage
Data profiling	Basic	Basic	Very good	Basic	Good	Good	Very good	Very good
Data parsing/ correction	Good	Very good	Very good	Good	Very good	Very good	Very good	Very good
Matching/ de-duplicate	Good	Very good	Very good	Good	Very good	Very good	Good	Very good
Enrichment		Very good	Very good		Very good		Very good	
Integration	Good	Good	Basic		Good	Very good	Basic	
Data monitoring			Good					
Data types	Customer Financial Inventory Materials data	Mutiple –mainly customer	Customer and product related data	Customer and non-customer	Customer and non-customer	Multiple	Multiple	Name and address; and non-name data
Real time support		Good	Good	Very good	Good	Good	Very good	Good
Database support		Good	Good	Basic	Very good	Good	Very good	Good
Enterprise Transactional Systems		Good	Very good				Very good	
Support for SOA	Basic		Good		Very good		Good	Very good
Integrated repository of rules and reference data	Basic				Very good	Good	Very good	
Define new rules							Basis	
Support for Metadata			Good			Good		Good
Geocoding		Very good		Good			Good	
Support for Unicode	Basic	Very good	Basic		Good		Good	
Client Server platform	Good	Basic	Good		Good		Very good	
User Interface		Intuitive	Graphical workflow GUI		Intuitive	Effective		Easy to use
Scalable		Good			Very good		Good	

In addition to the above mentioned data cleansing tools, we also investigated some more commercially available tools as well as tools developed as an outcome of academic research, like: AJAX (INRIA, France), Arktos (National Technical University, Athens), DataLever Enterprise Suite (DataLever Corporation), dn: Clean (Datanomic), Intelliclean (National University of Singapore), Porter's Wheel (University of California, Berkeley), and WinPure (WinPure Inc). These tools are not included in the comparison shown in Table 1, as they did not offer an exhaustive range of data cleansing features.

5.2 Comparison of Data Cleansing Tools

The data cleansing software tool 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 full spectrum of capabilities. It becomes convenient for organizations to deploy a single-vendor solution for enterprise-wide data quality requirement. Table 1 provides a comparison of eight major data cleansing tools we investigated in detail. The three levels of rating: 'basic', 'good' and 'very good' are assigned based on the degree of support for a feature by the vendor. Cells in a column are left blank where sufficient information about the respective features is not available from the vendor.

5.3 Analysis of Data Cleansing Tools

The initial impetus to commercial data quality tools was given by customer data management problem. The customer data (i.e. the name and address of customers) that support Customer Relationship Management (CRM) related activities is the most volatile field in corporate databases. The quality of this type of data quickly degenerates over time. Therefore, most of the pioneering data quality tools focused initially on cleaning up customer data only; even today majority of data quality functionality is aimed at this type of data. But at the same time the present day data quality tools have expanded well beyond such capabilities and the vendors are including other data domains like product data and financial data to their list. Master data management (MDM) initiatives are driving the product data whereas financial data is being driven by corporate accountability and regulatory pressures from governance initiatives like Sarbanes-Oxley, and HIPAA compliance.

Lots of applications are driven by today's powerful database technologies. The data cleansing tools are offering fast access to and from relational databases like: Oracle, SQL Server, DB2 and others. They are also supporting Windows ODBC connectivity, and traditional delimited files. In addition to conventional data quality functions, the tools provide a scope for connectivity to databases, integration with enterprise systems like Enterprise Resource Planning (ERP) and data warehousing tools like Extract transform Load (ETL).

Text processing systems worldwide are increasingly supporting Unicode to have a consistent way of encoding multilingual text and to exchange text files internationally. As enterprises expand their reach to global locales and clients, more and more companies are moving their business data to Unicode. Majority of vendors we reviewed are providing support for Unicode framework, allowing users to read and write data from a wide variety of Unicode and non-Unicode pages. Some vendors are supporting Geocoding and are able to enrich data by adding value to the captured data. This feature is very importance for assets that are installed in remote areas and for those assets that are in transit.

Adoption of Web services and service-oriented architecture (SOA) for achieving agility and ease of adapting to changing business requirement by organizations is at an increase. This is very well matched by the data cleansing tool vendors who are providing support for SOA. They are supporting industry Web Services standards like SOAP, XML, WSDL, UDDI, and HTTP through the SOA framework. SOA is an ideal implementation methodology for centralizing data quality processing across the enterprise and is quick to implement and easy to maintain. The tools should reach out with Web Services integration for popular IBM WebSphere, BEA WebLogic, and Microsoft .NET platforms.

6. MORE NEEDS TO BE DONE TO SUPPORT EAM DATA

The functionality of commercially available data cleansing software tools has improved over the years. The tools can adequately address the data quality problems of transaction based data that usually reside in tables of relational databases. But more needs to be done towards the improvement of specifications and feature of these tools enabling them to cater for the peculiar and unique data that are typical of asset management environment.

Some asset data may not be available in structured tables; it may reside on flat files. Though data cleansing tools have started to broaden their range by adding more data types like non-name product data and financial data under the structured data category, they are far from handling typical asset data that might include parts list and condition monitoring data under structured data types, and specifications, inspection reports, instructions etc. under unstructured data types.

The earlier versions of data cleansing tools worked on flat files in batch mode. The present data cleansing tools are supporting hybrid client/server architectures that allow validation, standardization, and matching done in real-time across a LAN or WAN (Lee et al, 1999). Support for real-time functions is very important in an asset management setting where control and monitoring data can be in real-time and streaming and it is expected to cleanse data before it is saved onto a database or data repository, or used for triggering or actuating some action linked to the status of data.

In asset management environment, annotations are often done by engineers and technicians during regular installation, operation and maintenance of assets and while carrying out modifications and improvements to the existing process or assets. This generates a lot of very useful metadata that calls for adequate handling. The future tools need to enhance their capabilities from syntactical to semantic. They not only need to merely recognise the structure of data but also understand the meaning of data through extensive use of metadata. The tools based on metadata-driven design will enable enterprises to move beyond defect inspection to solving data defects through root-cause analysis and building in data quality from the outset as a core function of any application design. The vendors like Trillium Software and Firstlogic have developed tools that make use of some data matching and cleansing business rules (Galhardas, 2001) but tools need to have centralised set of common business rules that can drive all the various data quality components like profiling, matching, cleansing, validation, standardization and enrichment.

7. CONCLUSIONS

By having an effective asset management in place, EAM organization can reduce total cost of ownership of their assets, get the most out of their assets, minimize the need for new assets, and continuously align capital assets with organizational needs. The organizations rely on a number of disparate information systems to improve operations and maintenance strategies for their engineering assets. Through these systems, they collect structured and unstructured data in various formats. Bad quality data in data repositories and lack of integration between diverse systems does not allow the organization to have a comprehensive view of all the data they own and severely impacts the quality of their decision making process.

Data cleansing software tools can provide help in improving data quality. But sadly, as shown above, data cleansing tools that are commercially available from vendors fall short of cleansing some nuances of engineering asset data. The vendors need to enhance their features to cater for the burgeoning need for cleansing asset management data.

This paper discusses how data cleansing tools can improve the quality of dirty data; though it should not be considered as the only way to improve the quality of data. Moreover data cleansing through tools should not be seen as a standalone activity. The real and more effective solution lies in establishing a well defined data quality methodology that controls the business process and does not allow the bad data to enter information systems in the first place.

8. ACKNOWLEDGMENTS

This research is conducted through the CRC for Integrated Engineering Assets Management (CIEAM) at the University of South Australia. The support of CIEAM partners is gratefully acknowledged. Any opinions, findings, and or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the CIEAM.

9. REFERENCES

Blanchard, B.S. (2006). System Engineering Management, 4th edition, New Jersey: John Wiley and Sons.

Dasu, T & Johnson, T. (2003). Exploratory Data Mining and Data Quality. New Jersey: John Wiley and Sons.

DFC (2006). SAS DataFlux Corporation Home page, URL www.dataflux.com/, Accessed 2 Apr 2006.

English, L.P. (1999). Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits, Willey & Sons.

Eerens, E. (2003). Business Driven Asset Management for Industrial & Infrastructure Assets. Le Clochard: Melbourne.

FLI (2006). FirstLogic Inc. Home page, URL www.firstlogic.com/, Accessed 3 Apr 2006.

Friedman, T. & Bitterer, A. (2006). Magic Quadrant for Data Quality Tools, Gartner Research Report.

G1S (2006). Group1 Software Home page, URL www.g1.com/, Accessed 5 Apr 2006.

Galhardas, H., Florescu, D., Shasha, D., Simon, E. & Saita, C.-A. (2001). Declarative data cleaning: Language, model, and algorithms. Proceedings of the 27th VLDB Conference, Roma, Italy.

Hernandez, M. A. & Stolfo, S. J. (1998). Real-world data is dirty: Data Cleansing and the Merge/Purge problem, Journal of Data Mining and Knowledge Discovery, 2(1), 9-37.

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

HTS (2006). Harte-Hanks Trillium Software Home page, URL http://www.trilliumsoftware.com/site/content/products/tss/index.asp, Accessed 2 Apr 2006.

IBM (2006). IBM Corporation Home page, URL http://www-306.ibm.com/software/data/integration/qualitystage/, Accessed 2 Apr 2006.

IMC (2006). Informatica Corporation Home page, URL http://www.informatica.com/solutions/data_quality/default.htm, Accessed 7 Apr 2006.

IPWEA. (2002). International Infrastructure Management Manual. Australia/New Zealand Edition. The Institute of Public Works Engineering Australia.

ISI (2006). Innovative Systems Inc. Home page, URL http://www.innovativesystems.com/, Accessed 6 Apr 2006.

Lee, M.L., Lu, H, Ling, T. W. & Ko, Y. T. (1999). Cleansing data for mining and warehousing. Proceedings of the 10th International Conference on Database and Expert Systems Applications, Florence, Italy.

Maletic, J.I. and Marcus, A. (2000). Data Cleansing: Beyond Integrity Analysis. Proceedings of the International Conference on Information Quality, MIT, Boston, USA.

Muller, H., Freitag, J., (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing, Technical Report HUB-IB-164, Humboldt University, Berlin, Germany.

Rahm, E & Do, H.H. (2000). Data Cleaning: Problems and Current Approaches, IEEE Data Engineering Bulletin 23 (4).

Snitkin, S. (2003). Collaborative Asset Lifecycle Management Vision and Strategies. Research Report. ARC Advisory Group.

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

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

Wang, R.Y., Storey, V.C. and Firth, C.P. (1995). A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering, 7(4), 623–640.