3-1-2010

Improving Data Quality in Conversion Projects: The Impact of Source Systems and Team Experience

Jeff Lucas
jslucas@crimson.ua.edu

David Hale

Joanne Hale

IMPROVING DATA QUALITY IN CONVERSION PROJECTS: THE IMPACT OF SOURCE SYSTEMS AND TEAM EXPERIENCE

Jeff Lucas
The University of Alabama
jslucas@crimson.ua.edu

David Hale, Ph.D.
The University of Alabama
dhale@cba.ua.edu

Joanne Hale, Ph.D.
The University of Alabama
jhale@cba.ua.edu

ABSTRACT

Poor data quality has been shown to have a serious impact on organization performance including increased operational cost and ineffective decision-making. In response to poor data, many organizations take on data cleansing projects as part of ERP and data warehouse implementations. These projects can be extremely difficult and produce less than desired results.

This study will examine the data cleanup efforts taken on by an organization specializing in implementing and maintaining benefits modules for an ERP system. In particular this study will build on research in traditional software development and examine the impact of the conversion and cleansing team's experience with the source systems, the target system and systems within a similar domain on the accuracy of data following the conversion and cleanup effort.

Keywords

Data Quality, Information Quality, Data Administration, Data Warehousing, ERP

INTRODUCTION / MOTIVATION

Poor data quality within enterprise systems can have a profound impact on organizational performance. Estimating a true cost of data errors within an organization can be very difficult, but the impact of poor data quality is easy to identify. "These impacts include customer dissatisfaction, increased operational cost, less effective decision-making and a reduced ability to make and execute strategy" (Redman, 1998).

The impact of poor quality data on customer satisfaction and increased operational costs can be seen in everyday retail and corporate transactions. An example would be a sales organization that instituted a new "salesman's briefcase" system. The system would track each sales person's client contacts, sales leads, customer contact information, etc. The system would also be used to drive commission pay. The data converted into to the new system was highly prone to data errors. In many cases accounts were tied to the wrong sales associate. Customer contacts were associated with the incorrect account. Sales within the organization slowed dramatically as the sales force focused on cleaning the data in the new system.

Given a large portion of the sales force's pay was tied to the accuracy of the data in the system, cleaning the data immediately became their top priority. The organization paid not only in terms of the effort needed to clean the data, but in lost sales and a dissatisfied customer base that was not being attended to by the sales force.

While poor data quality's impact on decision-making and the ability to make and execute strategy may be difficult to measure, it is also easy to recognize. According to one executive "we spend about half our (decision-making) time just arguing about whose data is better" (Redman, 1998). This phenomenon is particularly pronounced in data warehousing and Enterprise Resource Planning (ERP) projects and systems.

A data warehouse (or smaller-scale data mart) is a specially prepared repository of data created to support decision making. "Providing high quality data to decision makers is the reason for building a warehouse" (Wixom and Waton, 2001). As seen in the Wixom and Waton study (2001), a high level of data quality is associated with a high level of perceived net benefits. If the underlying data can not be trusted and often is conflicting, how can it be used to drive business decisions and strategy?

ERP systems provide “two major benefits that do not exist in non-integrated departmental systems: (1) a unified enterprise view of the business that encompasses all functions and departments; and (2) an enterprise database where all business transactions are entered, recorded, processed, monitored, and reported” (Umble et al., 2003).

This unification of data into a single database places a premium on the accuracy of the data. Inaccurate data introduced by one department within the organization will now have a domino effect on the entire organization (Umble, et. al., 2003). Correcting existing data errors in source systems during the conversion process becomes a critical, albeit difficult process. As a senior project manager at a firm specializing in EPP system implementations put it, “Even when an implementation goes really well and we nail the entire project, data is still a serious problem.”

With the underlying critical need for clean, accurate and usable data, organizations and researchers alike have spent more than a decade looking for better ways to define and maintain accurate data. The reader is directed to Lee et al. (2002) for an overview of related research. This prior research has focused primarily on the taxonomy of data errors, the process for data cleansing and maintenance and the tools used in the data cleansing process. Despite this prior work, data cleansing efforts tend to be very difficult and quite costly to organizations.

This study will focus on organizations efforts to clean data needed to support business functions, both transactions and decision making. In particular it will examine the typical process taken to “cleanse” data during a conversion and how task dependent metrics are impacted by the make-up and experience level of the data conversion and cleansing team.

LITERATURE REVIEW

Dimension of Data Quality

Much of the prior research to define data quality or information quality (IQ) has focused on defining the multiple dimensions of information quality. In general IQ dimensions can be grouped into four categories (Lee et al., 2002):

- Intrinsic IQ – Information has quality in its own right
- Contextual IQ – Information should be considered within the task at hand
- Representational and Accessibility IQ - Systems should provide access to information in a way that is easy to understand and manipulate.

Assessing Data Quality

It is important when determining the quality of data within an organization, to review the data through the lens of the user. “If stakeholders assess the quality of data as poor, their behavior will be influenced by this assessment” (Pipino et al., 2002). This study will adopt this user-centric focus and define data or information quality as “data that are fit for use by data consumers” (Wang and Strong, 1996).

Data quality assessments can be segregated into task-independent and task-dependent metrics. “Task-independent metrics reflect states of the data without the contextual knowledge of the application, and can be applied to any data set, regardless of the tasks at hand. Task dependent metrics, which include the organization’s business rules, company and government regulations, and constraints provided by the database administrator, are developed in specific application contexts” (Lee et. al., 2002).

Given the definition of information quality as data that are fit for use by the data consumer, task dependant metrics are vital. For example, an organization that chooses to implement an ERP Benefits module to automate retirement calculations could have a very different opinion of their data before and after the implementation:

- Prior to the ERP implementation many processes and functions are executed manually. For example when an employee initiates retirement, a retirement specialist from the benefits team is likely to complete the pension calculation manually. In order to complete this process the specialist would gather needed data from payroll, HR, etc. In this black box world of the manual process the user would have given a high quality data rating.
- Once the pension calculation is automated, the same data is needed in a standardized format within the ERP system. In the post-conversion world where retirements are initiated and processed on line, holding the process up while a retirement specialist tracks down needed data would likely lead to a frustrated employee and a retirement specialist that is not satisfied with the state of their data.

When assessing the state of data quality within an organization, both subjective and objective measures must also be considered (Pipino et. al, 2002). Subjective measures deal primarily with the data consumer's perception of the data quality. These subjective measures can be captured with the use of a questionnaire. As mentioned earlier, these subjective measures are important as they drive user behavior. The most straight forward objective measure is a simple ratio. "The simple ratio measures the ratio of desired outcomes to total outcomes" (Pipino et. al, 2002).

Data Complexity

When completing a data conversion, cleansing or integration project as seen in many data warehouse or ERP implementations, the cleansing team must determine the complexity of the data that will be received from the source systems. This research proposes that the following items drive this complexity:

- The total volume of data as measured by the number of records in the source systems,
- The number of data elements being integrated or cleansed,
- The number of sources of data that will be utilized and
- The number of non-automated sources of data.

In addition to the volume of data, data complexity of the source data is driven by the understandability and usability of the data storage system as well as redundancy of data across systems. If there is a large amount of redundancy of data in various source systems it will require additional data integration activities and increase the complexity of the source system data.

Data Quality Drivers

This research proposes that the primary drivers of target data accuracy in a data conversion and cleansing project are the data quality of the source systems and complexity of data received from the source systems. If an organization is implementing a new ERP system and conducting a cleansing project as part of the implementation the first step in the project would be to assess the current state of the data.

If there is only a single source of data, that is easy to access and understand, and the data within the source system is high quality, the target system will have high quality data. Unfortunately, clean data with low complexity is seldom the starting point of a data cleansing project. As the quality of the data declines or the complexity increases other factors, such as tooling, advanced analytical techniques and team experience, play an important role in the data cleansing project.

Tooling tends to be most effective improving task independent measures of data quality. There are several tools in the marketplace that can be used to cleanse specific domains such as name or address data or to complete duplicate record elimination (Rahm and Hai Do, 2000).

To improve task dependent measures of data quality advanced analytical techniques can play a vital role. There are many approaches to data analysis, but data profiling and data mining are often used. Data profiling focuses on instance analysis such as data type, length, value range, discrete values, variance, uniqueness, typical string values, etc. Data mining efforts includes clustering, summarization association discovery and sequence discovery (Rahm and Hai Do, 2000). The information gained from this analysis can be used to generate highly sophisticated mapping and cleansing rules.

Given the task dependent nature of data quality, it is proposed here that the cleansing team's knowledge, experience and ability would significantly influence the outcome of the cleansing project. A review of current software development literature reveals that experience within the same system leads to productivity enhancement for an individual programmer and within programming teams (Boh et al, 2007). This same study found that team productivity was enhanced not only by experience within the same system, but also within related and similar systems. In fact experience with related systems had a greater impact on productivity at the team level than did experience within the same system.

RESEARCH OBJECTIVE AND QUESTIONS

This study looks to add to the data cleansing literature by analyzing the impact of the cleansing team's experience on data quality. In particular, this study will build on the Boh et al. (2007) findings and determine the impact of team experience with the target system, the source system and similar systems on target data quality.

Building on this model, this study will test three primary hypotheses:

H_{1a}: Increased team experience with the source system will increase target data quality in data cleansing projects.

H_{1b}: The impact of team experience with the source system will be greater for highly complex data sets or source systems with extremely poor data quality.

H_{2a}: Increased team experience with the target system will increase target data quality in data cleansing projects.

H_{2b}: The impact of team experience with the target system will be greater for highly complex data sets or source systems with extremely poor data quality.

H_{3a}: Increased team domain knowledge will increase target data quality in data cleansing projects.

H_{3b}: The impact of team domain knowledge will be significantly greater for highly complex data sets or source systems with extremely poor data quality.

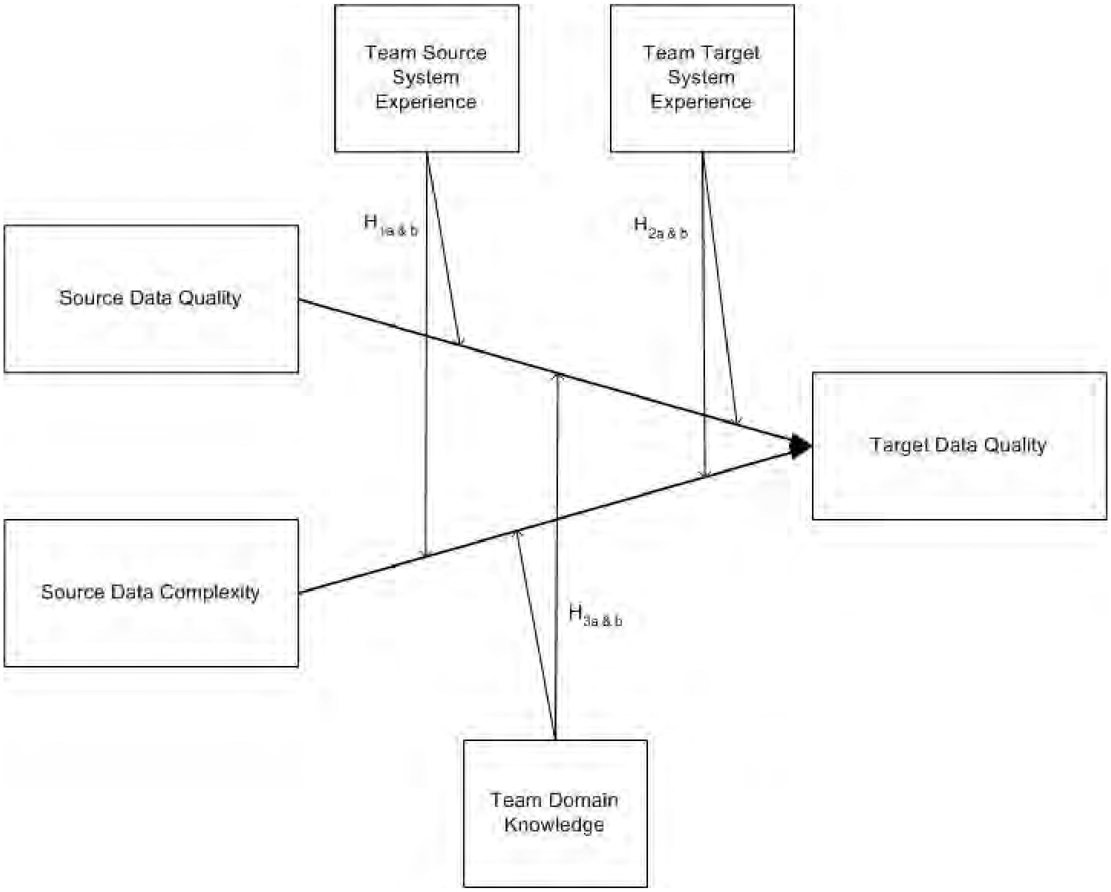

Figure 1. Data Quality Success Drivers and Moderators

METHODOLOGY

This study will employ a combination of methods in order to triangulate and confirm findings. The primary method employed will be a review of project archival data of nearly 60 conversions executed within a single firm in the past two years. This will include a review of the issue and defect tracking system for the conversion and cleansing project as well as a review of the defects created and tracked post conversion.

The archival review will also include the data conversion requirements documentation, conversion development specifications and conversion reconciliation reports. Through this archival data the research team will quantify the team experience, source data accuracy, source data complexity and target data accuracy. Surveys will be conducted to measure team experience and user perception of data quality. A more detailed explanation of each variable follows:

Source Data Accuracy

Source data accuracy will be measured using the conversion team's pre conversion profiling and edit reports. These reports detail the state of the data as received by the conversion team including the total number of records with missing or incomplete data, the number of records that fail to meet formatting standards as well as the number of records that do not pass context specific edits.

Each error will be categorized and given a severity rating from one to three based on the impact the error would have on ongoing processing if not corrected. A pre conversion data score will then be calculated by summing the severity rating of each error. This pre conversion data score will be assessed in conjunction with the total number of records to assign a source data accuracy rating of high, medium or low.

Data quality must also be measured through the eyes of the user. If the ultimate data user does not believe the data is of high quality, it will impact his use of the system (Pipino et. al., 2002). For this reason a survey instrument will also be employed to measure the conversion team's assessment of the source data quality. The conversion team's perception of the source data quality will be considered in the context of the quantitative data accuracy rating to determine the final source data accuracy rating.

Source Data Complexity

Source data complexity will be measured using the conversion data requirements, development specifications and team analysis meeting notes. The requirements documentation will detail the total number of sources, the number of sources that have overlapping or competing records as well as the structure and schema of the source data.

As with data accuracy, data complexity is also best examined through the lens of the user. If the data is complicated and difficult for the user to understand, the conversion will be more difficult to execute. The review of project archival data will be used in conjunction with the survey results to assign a source data complexity rating of high, medium or low.

Target Data Accuracy

Target data accuracy will be measured using three sources of data. First the team will review the number of records flagged for manual processing at conversion due to known data defects. This count will be considered in relation to the total number of records in the target data. Next the team will review the number of processes that are stopped or hit processing edits due to data in the first three months following conversion. Finally the issues and defects logged in the first three months following the conversion will be examined to identify data related defects. This information will be used to assign a target data accuracy rating of high, medium or low.

A survey instrument will also be employed to measure the ongoing team's assessment of the source data quality. The ongoing team is responsible for processing transactions using the conversion data and the ultimate user of the target data. The ongoing team's perception of the target data quality will be considered in the context of the quantitative data accuracy rating to determine the final target data accuracy rating.

Team Experience

A survey instrument will be used to elicit the number of month's experience each team member had with the source systems prior to the conversion as well as the number of months experience each had with the target system and other systems in the same domain. Follow-up interviews with the conversion team lead will be conducted in order to evaluate overall team proficiency with the source systems, target system and the conversion tool set in general.

RESEARCH STATUS

A review of literature pertaining to data accuracy, data complexity, data cleansing, data migration, data maintenance and experience based learning in software development is underway. The target research organization has been contacted and the terms and conditions for access to organizational data are being negotiated.

At the conference the authors will present a research in progress report on the initial findings of the study including insights gained during the initial phases of the field study, any additional drivers uncovered by the research team, etc.

REFERENCES

1. Ballou, D. and Pazer, H. (1985) Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems, *Management Science*, 31,2, 150-162,
2. Boh, W., Slaughter, S. and Espinosa, J. (2007) Learning from Experience in Software Development: A Multilevel Analysis, *Management Science*, 53, 8, 1315-1331.
3. Delone, H. and McLean, E. (1992) Information Systems Success: The Quest for the Dependent Variable, *Information Systems Research* 3, 1, 60-95.
4. Goodhue, D. (1995) Understanding User Evaluations of Information Systems, *Management Science*, 41, 12 1827-1844.
5. Jarke, M. and Vassiliou, Y. (1997) Data Warehouse Quality: A Review of the DWQ Project, *Proceedings of the Conference on Information Quality, Cambridge, MA* 299-313.
6. Lee, Y., Strong, D., Kahn, B. and Wang, R. (2002) AIMQ: a methodology for information quality assessment, *Information and Management*, 40, 133-146.
7. Pipino, L., Lee, Y. and Wang, R. (2002) Data Quality Assessment, *Communications of the ACM*, 45, 4, 211-218.
8. Rahm, E. and Hai Do, H. (2000) Data Cleaning: Problems and Current Approaches, IEEE Computer Society, 23, 4, 3-13.
9. Redman, T. (1998) The Impact of Poor data Quality on the Typical Enterprise: Poor Data Quality Has Far-Reaching Effects and Consequences, *Communications of the ACM*, 41, 2, 79-82.
10. Wand, Y. and Wang, R. (1996) Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, 40, 5, 103-110.
11. Wang, R. and Strong, D. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, 12, 4, 5-34.
12. Wixom, B. and Watson, H. (2001) An Empirical Investigation of the Factors Affecting Data Warehousing Success, *MIS Quarterly*, 25, 1, 17-41.
13. Umble, E., Haft, R., and Umble, M. (2003) Enterprise Resource Planning: Implementation Procedures and Critical Success Factors, *European Journal of Operational Research*, 146, 241-257.
14. Zmud, R. (1978) Concepts, Theories and Techniques: An Empirical Investigation of the Dimensionality of the Concept of Information, *Decision Sciences*, 9, 2, 187-195.