DATA QUALITY ISSUES IN INTEGRATED ENTERPRISE SYSTEMS

Diane M. Strong
Worcester Polytechnic Institute
dstrong@wpi.edu

Olga Volkoff
Simon Fraser University
ovolkoff@sfu.ca

Abstract: In this paper, we explore the effects of Enterprise Systems, such as SAP, on the quality of organizational data. Our findings are based on a longitudinal study of an ES implementation at multiple sites in one global organization. We organize the results in terms of how the ES implementation changed Intrinsic data quality, Representational data quality, Contextual data quality, and Accessibility data quality (Wang and Strong, 1996). Our findings indicate improved data quality in all these categories, but the improvements were at a global level. At the local level of a particular functional area, the quality of data, in terms of its "fitness for use", might decrease.

Key Words: Data Quality, Enterprise Systems, ERP Systems

INTRODUCTION

Enterprise systems (ES) have become common in large and medium-sized organizations. These systems, with their common integrated database that serves the needs of most organizational users, potentially have a large effect on the quality of data available to organizational data users. In this paper, we explore the effects of an ES on data quality using data gathered during a longitudinal study of an ES implementation process.

BACKGROUND LITERATURE

To consider the effects on data quality (DQ) of installing and using an ES, we first briefly discuss the definition of DQ and the basic characteristics of an ES.

Data Quality Definition and Dimensions

We define data quality following the "fitness for use" standard in the quality literature. High quality data are data that are fit for use by those who use data (Wang and Strong, 1996). To understand what "fitness for use" means, the dimensions of data quality have been studied and grouped into four overall categories (Wang and Strong, 1996), as shown in Table 1, and these dimensions have been validated as useful for understanding data quality issues in organizations (Strong, Lee, and Wang, 1997).

To summarize Table 1, *Intrinsic DQ* denotes that data have quality in their own right. *Representational* and *Accessibility DQ* emphasize the importance of the information systems that store and provide access to these data; that is, the system must present data so that they are interpretable, easy to understand and

manipulate, and represented concisely and consistently and the system must be accessible but secure. *Contextual DQ* highlights the requirement that DQ must be considered within the context of the task at hand; that is, data must be relevant, timely, complete, and appropriate in terms of amount so as to add value (Lee, Strong, Kahn, and Wang, 2002). We use these DQ categories and dimensions as we investigate how ES's affect DQ.

Table 1: Data Quality Dimensions (from Wang and Strong, 1996)

DQ Category	DQ Dimension
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation
Contextual DQ	Relevancy, Value-added, Timeliness, Completeness, Amount of data
Accessibility DQ	Accessibility, Access security

Enterprise Systems

Enterprise Systems are "commercial software packages that enable the integration of transactions-oriented data and business processes throughout an organization" (Markus & Tanis, 2000, p. 176). These commercial packages, provided by vendors such as SAP, PeopleSoft, and Oracle, are large, expensive, and hard to implement. An objective of implementing an ES, or enterprise resource planning (ERP) system, is to provide seamless support and integration across the full range of business processes, uniting functional islands and making their data visible across the organization in real time. To reap these benefits, however, organizational units must be willing to develop common definitions for data and processes across the organization (Strong and Volkoff, 2004).

METHOD

Because our overall research goal was to understand ES effects in organizations, we chose to employ intensive methods of prolonged field observation coupled with interviews and document analysis. Our longitudinal approach entailed weekly site visits over a three-year period starting in August 2000, resulting in approximately 150 site visits.

Our primary research site was ACRO, a multinational producer of high-precision industrial equipment. The company's headquarters and twenty operating locations are located in the U.S., with additional plants and after market service facilities spread over eleven other countries. In 1998 ACRO embarked on a plan to implement ES software from SAP, configured as a single instance, throughout the company. We selected ACRO as our primary research site because it was undertaking a major ES implementation within which a number of distinct but comprehensive implementation projects were planned. The implementation projects we observed included (1) the manufacturing and assembly operations, (2) two service and repair sites, (3) the product research and development organizations, and (4) contract management.

We conducted recurring interviews with power users, other users, and managers at multiple points in time during the ES implementation process. For each implementation project, we conducted interviews before go-live to understand operations in the legacy system environment, a few months after go-live to observe short-term effects, and a year after go-live to observe longer-term effects. The resulting data included transcripts from 72 interviews, some with multiple participants, for a total of 92 person interviews, and extensive field notes.

These interview data were all coded and analyzed following grounded theory techniques (Glaser and Strass, 1967). As data were collected they were coded. During our open coding process, "data quality" emerged as a code, that is, it was one of the observed changes enabled by the ES. Open coding proceeded until no further codes were added to our list of codes. For this paper, a second coding pass, axial coding (Strauss and Corbin, 1998), focused on coding for subcategories of data quality. For this, we used the previously developed categories and dimensions of data quality (Wang and Strong, 1996) as our family of codes (Glaser, 1978). Thus, we treated the data quality categories as a theoretical sensitizing device or lens (Klein and Myers, 1999) for viewing the effects of an ES on data quality. The findings described below were produced as a result of these analyses.

FINDINGS

We present our findings about how ES's affect DQ below, organized in terms of the four categories of DQ. For each category, we present the data quality issues we observed with examples from our field data. Our findings are summarized in Table 2.

Intrinsic DQ: Need for Higher Accuracy

From our data, we observed three ways in which the ES "demanded" data of higher accuracy than in the legacy systems the ES replaced. First, there is more reliance on data and less ability to work around the system. For example, picking a part to send to the shop floor requires a demand for that part in the ES and accurate inventory values. As noted by one inventory planner,

> "You have to have the proper inventory levels. What you say you have you should have and what you don't have you should not have. It's that simple."

The transactions for issuing parts check which demand a part issuance meets and whether inventory is available. Overall, the ES works in a way that requires accurate data.

Second, because the ES is an integrated system, the accuracy needed for any data item is the highest accuracy level needed of anyone using that data item. For example, in the legacy world when design engineering and the manufacturing floor used different, unintegrated systems, engineering could specify an engineering change order to the accuracy required in the engineering system and send a paper memo to manufacturing about the change. In the integrated ES, engineering is connected to manufacturing and must specify an engineering change with the accuracy and level of detail required by manufacturing, which requires more data of higher accuracy.

Third, the ES is primarily an operational level transaction processing system. Yet, it also stores plans, e.g., production plans and budgets, and other estimates which it treats as it would any other operational data. Plans, however, are never as accurate as records of completed transactions. The ES treats these estimates as though they were accurate, thus, requiring higher accuracy in future data than was required in legacy systems. In the legacy environment, planning data, e.g., production plans and demand forecasts, were kept in separate systems, and while they were regularly compared to actual events, they were not an integrated part of the operational system used for daily decision-making.

Representational DQ: Common Integrated Data Definitions

Because it is integrated, an ES places different demands than the legacy systems on the representation of data (Volkoff, Strong, and Elmes, 2005). Specifically, an ES is built on an integrated database. This requires different groups to develop common data definitions. The following two examples were representational DQ problems observed at our field site. Finance and manufacturing differed in their definitions of part location, especially for sub-components in the process of being manufactured. This was not a problem when these groups used separate legacy systems, but in the common integrated database one definition was needed. The one chosen determined part location as a by-product of

transactions about job completion and labor hour charging. This definition was preferred by finance because it captures the group responsible for the part. Manufacturing, on the other hand, preferred physical location of the part, because as one manufacturing manager noted:

> "You can be looking for a part in one area and it's still residing in the previous area, which makes it difficult."

Manufacturing's definition would have required a separate transaction to be entered when a part or sub-component is moved to a new location. By using finance's definition, location is determined as a by-product of recording job completion and coincides with the definition of responsibility, but requires extra work for manufacturing when it is trying to locate in-process parts.

Table 2: Summary of Findings

DQ Category	Increased Fitness for Use	Decreased Fitness for Use
Intrinsic DQ	The ES demands and enables more accurate data – they are needed for transactions to work, which is needed because workarounds are difficult.	
	The ES "requires" each data item to be as accurate as the highest accuracy needs of any group.	Groups specifying data have to specify more data of higher accuracy than they have or need.
	The ES stores planning data like it stores operational data, thus requiring higher accuracy for planning data.	Because plans are not accurate and must be changed as better data arrive, users must frequently engage in the difficult processes of changing data in the ES.
Representational DQ	Sharing data through a shared database relies on common data definitions, e.g., location of parts, serial numbers.	A common data definition may provide a solution that favors one group over another (finance over manufacturing or manufacturing over repair).
		The ES uses abstract, hard to remember numbers rather than part numbers.
Contextual DQ	Common data definitions promote the sharing of data globally. Data become relevant in multiple contexts resulting in more effective use of data.	Local data definitions tailor data to the needs of local organizations, thus promoting efficient use of data.
		Whether data are of good quality depends on the local context. What is sufficiently accurate or timely for one group may not be so for another group.
Accessibility DQ	Through the common database, the ES promotes data accessibility and visibility by all.	To provide security, authorizations in the ES limit accessibility to transactions and associated data to only the few who need to execute each transaction.

A second example is serial numbers. The common, integrated part master file has one indicator for whether a serial number is tracked for a part. If it is, then every time a serialized part is handled, e.g., picked from the warehouse or removed from a product being serviced, the ES asks the user to enter the serial number. Unfortunately, the parts that are serialized by the R&D, manufacturing, and service organizations differ. For example, manufacturing works at the part number level and serializes parts whereas the repair organizations work at the level of removal components and serializes at the component level. R&D serializes to track wear and failures of new parts during development. A single serial number indicator was not sufficient for ACRO. It temporarily chose to serialize according to manufacturing requirements, and was searching for ways to handle the R&D and service organizations' differing serialization needs. These two examples illustrate the inherent difficulties of using common data definitions to satisfy multiple users with differing requirements.

A second representational DQ issue we observed was the move towards more abstraction in the data in the ES, similar to the abstraction noted by Zuboff (1988). For example, in its legacy system, manufacturing used part numbers as the primary identifier of parts. The ES includes part numbers, but the data must often be searched by abstract numbers, i.e., batch numbers and service request numbers. Manufacturing workers, who are familiar with part numbers and used to remembering and using them in their systems, find this level of abstraction difficult and frustrating.

Contextual DQ: From Local to Global Context

The examples above of the need for higher quality data and for common data representations shift the context for data quality from a local, functional context, e.g., manufacturing or engineering, to a global shared context. Thus, data definitions and levels of accuracy that were relevant and of value, i.e., fit for use, in a local context were changed to fit the global context. The net results are data of relevance and value in multiple contexts, but of less relevance and value in some local contexts. To meet the needs of all users, those collecting or generating data, e.g., design engineering specification of engineering changes, had to collect or generate additional data as compared to the legacy environment to meet the completeness requirements of all users. For example, one inventory purchaser noted that:

> "You must have all this data prior to issuing a purchase order, when in reality, you really may not need that."

Meeting the needs of multiple user groups in one system has always been difficult (Markus, 1983, Strong and Miller 1995), and the integrated nature of an ES does little to alleviate these difficulties.

We observed one general situation in which local needs overrode global needs. That situation is an example of the accuracy vs. timeliness tradeoff (Ballou and Pazer, 1995). Because an ES is integrated, any data entered becomes available to all users. A side-effect of this is that changing data in the ES is difficult, since follow-on processing may have occurred. For estimated data that are not yet at their final level of accuracy, the cost of making changes led those who generated and entered these data to withhold them as long as possible. For example, customer orders for service when a product arrives for service were always somewhat inaccurate because the extent of needed repairs is unknown until the product is disassembled. Changing service orders is cumbersome because changes are included as addendums that might not be read by service technicians. Thus, those who entered and managed service orders delayed their entry into the ES, and kept private records, as long as possible. The side-effect was data that were not as timely as possible for those needing to generate part replacement orders to cover service needs and those forecasting manpower. This example illustrates the general case that data entry may be delayed to improve accuracy of the entered data, which provided less timely data to others who relied on that data.

Accessibility DQ: Increased Accessibility and Security

Because of the common, integrated database, an ES generally increases accessibility and visibility to organizational data (Elmes, Strong, and Volkoff, 2005). For example, local inventory planners could see not only the parts available in their local warehouse as with their legacy systems, but also they could see and draw on inventory in any production warehouse in the company. As one inventory planner noted:

> "Especially when you have a critical shortage, you can opt to look at the world wide inventory. Whereas before, we didn't have access to information like that … at a flick of a button you can grab that information."

Accounting and customer service representatives could see and monitor product progress on the shop floor. Due to this visibility, customer service representatives provided better service because they could see all transactions related to a customer order, including financial transactions that they could not see in their legacy system. Those in accounting could run daily trial balances by product and thus were immediately aware of any budget deviations. Line items flowed to the customer invoice as they occurred rather than at the end when the product was ready to ship. This level of accessibility and visibility of all by all enables better immediate control over operations (Elmes, Strong, and Volkoff, 2005).

While the ES provided increased visibility, it also provided increased security by limiting the transactions that could be performed by a given role. In the legacy system, workers with access to the system could often perform any transaction in the system. Thus, it was possible for a shop floor manager to generate a requirement, request a part to fill that requirement, and request that the part be immediately picked from the warehouse and delivered to the manufacturing floor. In the ES, no role has such extensive authority. Authority to perform transactions is carefully specified by role, so that a worker cannot perform transactions across multiple roles.

DISCUSSION

Examining the examples of changes in data quality within the observational data from our longitudinal study of an ES implementation presents some interesting issues. As summarized in Table 2, the ES enables, encourages, and even demands higher quality data. It demands higher accuracy and common data representations and provides increased accessibility and security. Thus, we might conclude that an ES promotes higher quality data and increases the fitness for use of data.

On the other hand, we might also conclude that the ES reduces fitness for use of some data in some local contexts. The issue is that the ES focuses on data quality in the global context rather than individual local contexts. Thus, data may be too accurate for some local contexts, requiring the gathering of more detailed and accurate data than needed in that context. The data may be too abstract or not defined in a way that best suits the local context. In addition, we observed accuracy vs. timeliness tradeoffs in which the local context prefers to enter data only when they are more accurate, resulting in reduced timeliness for users in other contexts.

To begin to develop a theoretical understanding of how an ES affects data quality, we need to analyze these conflicting findings. As a first step toward a theoretical understanding, we investigate whether the conflicting observations represent contingencies or paradoxes. For example, conflicting observations may represent different situations, some of which may increase data quality and some of which may decrease data quality. In this case, a theoretical understanding of the observed data quality effects involves specifying which conditions lead to increased data quality and which to decreased data quality. On the other hand, the conflicting findings we observed may not be separable into different conditions, but may result from the same condition. That is, there may be some characteristic of an ES that leads to both increased and decreased data quality simultaneously.

Our preliminary analysis supports the latter situation, namely that data integration leads to both higher data quality and to lower data quality. The primary mechanism by which an ES affects data quality is the integrated database. That is, the integration of data so that data items can be used by multiple groups is the driver for common data definitions and formats (Volkoff, Strong, and Elmes, 2005) and the resulting effects on data quality. Other mechanisms in the ES also contribute to the data quality effects. For example, the increased process structure and the associated discipline required of users of the ES (Elmes, Strong, and Volkoff, 2005) reduces the ability of users to work around the system or to otherwise act in ways inconsistent with the ES, thus increasing conformity with the "demands" of the ES, e.g., for increased quality of data. Identifying these mechanisms provides an explanation for the process by which the conflicting observations arise. For example, common data definitions can both increase and decrease data quality. Further analysis, however, might indicate that for some data, common definitions increase data quality, but for others, data quality decreases. Such a result would support a contingency analysis.

CONCLUSION

In summary, the ES does lead to higher quality data along many dimensions. This higher global quality can, however, contribute to lower fitness for use locally. Thus, paradoxically, we observed higher data quality, which in turn actually produced lower data quality for some users. Further analysis is needed to assess the extent to which our conflicting results can be explained by a more detailed modeling of when and how enterprise systems increase and decrease data quality.

ACKNOWLEDGEMENTS

This research was supported in part by a grant from the National Science Foundation SES-0114954. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the organization we studied for generously providing open access to observe their implementation processes, attend meetings, and interview employees, at a time when they were extremely busy. We also thank the many employees who willingly talked about the joys and frustrations of being part of an ES implementation.

REFERENCES

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

[2] Elmes, M. B., D. M. Strong and O. Volkoff, Panoptic Empowerment and Reflective Conformity in Enterprise Systems-Enabled Organizations, *Information and Organization 15*(1), January 2005, pp. 1-37.

[3] Glaser, B. G., *Theoretical Sensitivity*, Mill Valley, CA: Sociology Press, 1978.

[4] Glaser, B. G. and A. Strauss, *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine, 1967.

[5] Klein, H. K. and M. D. Myers, A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems, *MIS Quarterly* 23(1) 1999, pp. 67-93.

[6] Lee, Y. W., D. M. Strong, B. K. Kahn and R. Y. Wang. AIMQ: A Methodology for Information Quality Assessment, *Information & Management 40*(2), December 2002, pp. 133-146.

[7] Markus, M. L. Power, Politics and MIS Implementation, *Communications of the ACM 26*(6), June 1983, pp. 430-444.

[8] Markus, M. L. and C. Tanis, The Enterprise Systems experience - From adoption to success. In Zmud, R. W., Ed, *Framing the Domains of IT Management: Projecting the Future Through the Past*, Cincinnati, OH: Pinnaflex Educational Resources, 2000.

[9] Strauss, A. and J. Corbin, *Basics of Qualitative Research*, Second Edition, Thousand Oaks, CA: Sage Publications, 1998.

[10] Strong, D. M., Y. W. Lee and R. Y. Wang, Data Quality in Context, *Communications of the ACM 40*(5), May 1997, pp. 103-110.

[11] Strong, D. M. and S. M. Miller, Exceptions and Exception Handling in Computerized Information Processes, *ACM Transactions on Information Systems 13*(2), April 1995, pp. 206-233.

[12] Strong, D. M. and O. Volkoff, A Roadmap for Enterprise Systems Implementation, *Computer 37*(6), June 2004, pp. 22-29.

[13] Volkoff, O., D. M. Strong and M. B. Elmes, Understanding Integration Effects of Enterprise Systems, *European Journal of Information Systems 12*(2), June 2005, pp. 110-120.

[14] Wang, R. Y. and D. M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems 12*(4), Spring 1996, pp. 5-34.

[15] Zuboff, S. *In the Age of the Smart Machine: The Future of Work and Power*, New York, NY: Basic Books, 1988.