December 1999

Teaching Data Quality Concepts Through Case Studies

M. Pam Neely
Center for Technology in Government, SUNY Albany

Theresa Pardo
Center for Technology in Government, SUNY Albany

Teaching Data Quality Concepts Through Case Studies

M. Pamela Neely, Pneely@ctg.albany.edu
Theresa A. Pardo, Tpardo@ctg.albany.edu
Center for Technology in Government, SUNY Albany

Overview

It is estimated that as much as 75% of the effort spent on building a data warehouse can be attributed to back-end issues, such as readying the data and transporting it into the data warehouse. Data quality tools are becoming an increasingly important resource in preparing the data for the warehouse, thus enhancing the usability of the warehouse. This tutorial, based on current research in the field, will focus on a methodology for managing data quality issues. The tutorial will present a framework for identifying data quality issues and making sense of the data quality tools marketplace. A case study approach will be used. The methodology presented is applicable both as a tool to teach about data quality issues and as a tool to support practitioners as they seek mechanisms to facilitate the management of data, yet ensure appropriate data quality.

Background

Data warehousing is emerging as the cornerstone of information infrastructures in organizations. It is imperative that the issue of data quality be addressed if the data warehouse is to prove beneficial to an organization. Corporations, government agencies, and not-for-profit groups are all overwhelmed with enormous amounts of data. The desire to use this data as a resource for the organization has hastened the move towards data warehouses. This information has the potential to be used by an organization to generate greater understanding of their customers, processes, and the organization itself.

Attention to data quality is a critical issue in all areas of information resources management. An article in the Wall Street Journal (7/13/98) relates the domino effect that occurs when erroneous information is typed into a central database. A new airport in Hong Kong suffered catastrophic problems in baggage handling, flight information, and cargo transfer. The ramifications of the dirty data were felt throughout the airport. Flights took off without luggage, airport officials tracked flights with plastic pieces on magnetic boards, and airlines called confused ground staff on cellular phones to let them know where even more confused passengers could find their planes (Arnold, 1998). The new airport had been depending on the central database to be accurate. When it wasn't, the airport paid the price in terms of customer satisfaction and trust.

The steps for building a data warehouse or repository are well understood. The data flows from one or more source databases into an intermediate staging area, and finally into the data warehouse or repository (see Figure 1). At each step there are data quality tools available to massage and transform the data, thus enhancing the usability of the data once it resides in the data warehouse.

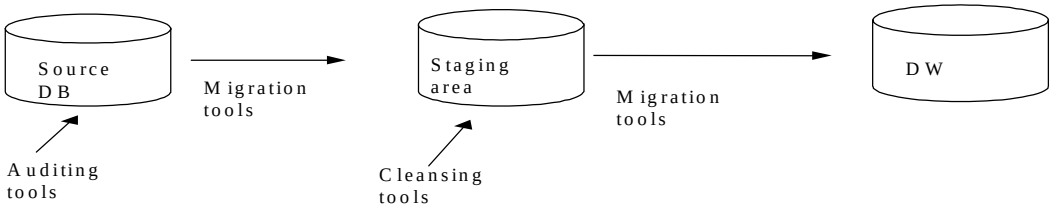

Figure 1- Data Flow and Data Quality Tools

Tutorial Focus

The objective of the tutorial is to increase the participants understanding of data quality issues and to share a framework for analyzing and addressing these issues. The focus of the tutorial will be on a framework that was developed to support the identification of data quality issues and the selection of tools for addressing those issues in several projects conducted at the New York State Center for Technology in Government. A case study approach and small group activity will be used. The case is designed to elicit discussion of data problems and to support identification of the related features of available tools to address those problems. The case study will focus on a large social services agency that engaged in the implementation of a data warehouse.

A primary component of the framework is the matrix, an excerpt of which is included at the end of the paper as Table 1, which facilitates the analysis of the data quality tools marketplace. The matrix supports the mapping of data problems to features of data quality tools. The framework presented in the tutorial allows the identification of data problems and guides the data warehouse developer to features of the tools that will address these problems. Proper analysis of data quality problems is essential to ensuring that the correct tool is chosen to automate the process of data cleansing. Discussion of the case will enable participants to evaluate data quality issues on all levels with important feedback from each other. Use of the matrix will allow the entire process to proceed more efficiently.

.

Expected Audience

This tutorial is aimed at educators, students, and practitioners. The case study approach, particularly with the active participation of a diverse group of participants, encourages interaction among the various groups, providing a richer experience for all involved. It is expected that the tutorial will run 1 ½ hours including discussion of data quality issues, introduction to the case, and small group activity followed by full group discussion and wrap-up

Table 1- An Excerpt from the Tool

"Mapping Data Problems to Features of Data Quality Tools"

Questions to be asked	Features	Tools
Auditing Tools		
Is your data complete and valid?	Data examination- determines quality of data, patterns within it, and number of different fields used	WizSoft- WizRule Vality- Integrity
Does your data comply to your business rules? (Do you have missing values, illegal values, inconsistent values, invalid relationships?)	Compare to business rules and assess data for consistency and completeness against rules	Prism Solutions, Inc.- Prism Quality Manager WizSoft - WizRule Vality- Integrity
Are you using sources with unknown business rules?	Data reengineering- examining the data to determine what the business rules are	WizSoft – WizRule Vality- Integrity

Biographies

Theresa Pardo is the Project Director at the Center for Technology in Government located at the University at Albany, State University of New York. The Center was created by New York State in 1993 to serve as a research & demonstration resource for New York's state and local governments. Through partnership projects that link public, private, and academic resources, the Center pursues new ways of applying computing and communications technologies to the practical problems of information management and service delivery in the public sector. The Center is a 1995 *Innovations in American Government Award* winner. As Project Director, Dr. Pardo works with a variety of government and corporate partners to manage projects which increase productivity, reduce costs, increase coordination, and enhance the quality of government operations and public services. Before coming to the Center in 1994, Dr. Pardo was the Director of Academic Computing Services at Siena College and the Assistant Director for Academic Computing at Union College. She holds a Ph.D. in Information Science from the Rockefeller College of Public Affairs and Policy at the University at Albany, State University of New York.

M. Pamela Neely is a graduate assistant at the Center for Technology in Government, University at Albany, State University of New York. She is researching data quality issues associated with the *Using Information in Government* project at CTG. In addition to her work at CTG, Pam is a lecturer in information science at the University at Albany and has been an adjunct instructor in computer science and accounting at Marist College. A certified public accountant for more than 20 years, Pam operates her own accounting firm. She earned a B.S. in environmental studies from SUNY Buffalo and an M.S. in accounting from the University of Colorado. Pam is a doctoral candidate in the Information Science Program at the University at Albany.

Selected References

Aragon, L. (1998). "Down With Dirt", PCWeek, February 9, 1998

Arnold, W. (1998). "Human Error Causes System Glitches the Embarrass New Asisan Airports", Wall Street Journal Interactive Edition. http://interactive.wsj.com.

Atre, S. (1998). "Rules for Data Cleansing", ComputerWorld.

English, L. (1998). "Data Quality: Meeting Customer Needs", Pitney Bowes white paper

English, L., (1996). "Help for Data Quality Problems", InformationWeek, October 7, 1996, pp. 53

Greenfield, L. (1998). "Data Cleaning, Extraction and Loading Tools" http://pwp.starnetinc.com/larryg/clean.html.

Haggerty, N. (1998). "Toxic Data", DM Review Magazine, June 1998

Horowitz, A. (1998). "Ensuring the Integrity of Your Data", Beyond Computing, May 1998

i.d. Centric (1998). "Customer Data Quality: Building the Foundation for a One-to-One Customer Relationship", i.d. Centric white paper

Kimball, R. (1996). "Dealing with Dirty Data" DBMS Online.

Moss, L. (1998). "Data Cleansing: A Dichotomy of Data Warehousing?", DM Review Magazine, February 1998

O'Neill, P. (1998). "It's a Dirty Job: Cleaning Data in the Warehouse", Gartner Group, January 12, 1998

Strange, K. (1997). "A Taxonomy of Data Quality", Gartner Group, May 29, 1997

Watterson, K. (1998). "Dirty Data, Dumb Decisions", DM Review Magazine, March 1998

Williams, J. (1997). "Tools for Traveling Data" DBMS Online, June 1997