AMCIS 2002 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2002

DATA QUALITY KNOWLEDGE MANAGEMENT: A TOOL FOR THE COLLECTION AND ORGANIZATION OF METADATA IN A DATA WAREHOUSE

M. Pam Neely
Rochester Institute of Technology

DATA QUALITY KNOWLEDGE MANAGEMENT: A TOOL FOR THE COLLECTION AND ORGANIZATION OF METADATA IN A DATA WAREHOUSE

M. Pamela Neely
Rochester Institute of Technology
mpnbbu@rit.edu

Abstract

This paper describes a relational database tool, the Data Quality Knowledge Management (DQKM), which captures and organizes the metadata associated with a data warehouse project. It builds on the concept of fitness for use by describing a measurement technique for subjectively assigning a measure to a data field based on the use and quality dimension of the data within the data warehouse. This measurement can then be compared to some minimum criteria, below which it is not cost effective to enhance the quality of the data. This tool can be used to make resource allocation decisions and get the greatest benefit for the cost in utilizing the scarce resources available to enhance source data for a data warehouse.

Keywords: Data quality, data warehouse, knowledge management, metadata, relational database

Introduction

Data quality is a critical success factor to many activities of the information age, including the development and operation of a data warehouse (Wixom and Watson, 2001). The issue of data quality is recognized in the development of the warehouse; however, there is no formal methodological approach to dealing with the quality issues. Additionally, although it is recognized that the quality of the metadata is essential to the quality of the data (Iverson, 2001), no systematic approach has been developed for the collection and analysis of metadata associated with a data warehouse project.

This research was targeted at the development of a relational database tool to improve the process of making data quality decisions in the context of the creation of a data warehouse. This tool, the Data Quality Knowledge Management (DQKM) tool, provides a method for effectively capturing and managing metadata in the complex environment that results when integrating multiple data sources.

Background information on data quality in the data warehouse will be presented. The remainder of the paper will concern itself with the DQKM, and the use of this tool in the collection and analysis of metadata for a data warehouse project.

Data Quality and the Data Warehouse

Data quality at the source is not the same as data quality in the warehouse. Data from the source is used for multiple purposes by multiple users with different needs for data quality and in ways that were not envisioned when the data was originally collected at the source. Two factors affect the quality of data in a data warehouse: quality dimension and use. The first factor is the dimension of quality being considered. Data quality is multi-dimensional (Wang and Strong, 1996). These dimensions include accuracy, timeliness, completeness, accessibility and reputation, among others (Wang and Strong, 1996). There can be trade-offs on these dimensions. The data may be accurate, but not timely. It may be complete, at the expense of concise representation. Another factor affecting the quality of data in the warehouse is how the data is to be used in the warehouse, or fitness for use (Ballou and Tayi, 1989; Redman, 1998; Strong, et al., 1997). Data whose quality is appropriate for one use may not be sufficient for another. A warehouse should be designed to address specific needs and answer specific questions. The quality of the data is dependent on how the data is used to address these needs and questions. Once the quality dimensions and uses have been defined, a subjective (or occasionally objective) measurement can be assigned to the data field. This measurement is a measure of the current quality of the data, as it would be used in the warehouse (as opposed to the quality at the source).

For example, a data warehouse designed to support decisions on services provided to homeless individuals will need to collect data on the clients in the system. The use of the data would be client demographics. A quality dimension to be considered may be timeliness. The data field that would be used is gender. In order to define data quality in the warehouse we must define this relationship among the data field (gender), the dimension (timeliness), and the use (client demographics). As stated earlier, the measurement is frequently subjective. Although this is a limitation, the measurement of data quality is an open issue and is an area of additional research (Ballou and Tayi, 1999). The warehouse development team would then use this measurement to make decisions regarding the allocation of resources for data quality projects. In order to assess the quality of the data, we need to understand the metadata and its implications within the data warehouse development project.

How do we identify the quality problems? In order to identify the problems we must become very familiar with the data and in order to fully understand the data, we need to know the metadata. However, the metadata for a data warehouse is found in many source documents and individuals. The amount of documentary evidence may be overwhelming for a data warehouse with numerous source databases. In addition, valuable metadata is carried in the minds of the data providers. The challenge is to organize the metadata into a format that can be used to identify the quality problems in the warehouse.

Commercial tools that are currently available do not completely integrate the technical specifications of the data with the business-oriented needs of the metadata. The DQKM is a decision support tool as well as a knowledge management tool. It provides a method for organizing the metadata in such a way that relationships among the various sources of data can be drawn. A key advantage of the DQKM is that it is a relational database, thus it is a queriable tool. The amount of metadata associated with a data warehouse can quickly become overwhelming to process. Harnessing the power of a relational database allows the warehouse development team to process the metadata in a more efficient and effective manner. Once populated, it can allow the data warehouse development team to make informed decisions on the benefits and costs associated with cleaning up specific data fields. In addition, the DQKM can be used to support decisions on how the data should be cleaned, manually or automatically. As a knowledge management tool, it captures the experience of the data providers with the data. When these individuals leave the organization, their expertise does not leave with them. Finally, novices as well as experts can use the DQKM. This results in an opportunity to shift some of the cost of development and decision-making from expensive experts to less expensive novices.

Data warehouse projects have finite resources, and the funds for assuring that the data are of high quality is some limited portion of those resources. So how do we decide where to invest these funds to return the greatest benefit? Ballou and Tayi (1989) propose a methodology for allocating resources for data quality enhancement. Their approach is to use an integer programming model for allocation of the total available resources based on various costs, the stored data error rate, and the effectiveness of the cleaning efforts. Increased awareness and sensitivity to data maintenance issues, which provide ancillary benefits in Ballou and Tayi's paper, becomes the focus of the research described in this paper. Using a systematic approach to evaluate the source data, we can assign a measure to the data fields that inform us of the importance of the data in the context of the warehouse. These measures can then be compared to a set of criteria and priorities. The resulting prioritized fields will then be considered for data quality enhancement.

Research Method

The Data Quality Knowledge Management (DQKM) tool was developed as part of a larger research project. This multi-method research involved the qualitative analysis of ten interviews conducted with data warehouse developers, as well as testing of a preliminary tool, an extension of a standard data dictionary (Neely, 2001). From this data, a framework was developed for the analysis of source data quality prior to the migration to a data warehouse. A significant focus of the framework is the collection and integration of the metadata associated with the data warehouse project. In order to implement this process, the DQKM was proposed. In an exploratory study phase of the research, the DQKM was populated with metadata drawn from a case study presented at AMCIS 1999 (Neely and Pardo, 1999). The results of the study indicate that there is potential for the tool to be used in making resource allocation decisions regarding the data quality projects associated with a data warehouse. This paper will describe the DQKM and the exploratory study, as well as some of the results from the study.

Data Quality Knowledge Management (DQKM)

As indicated earlier, the DQKM is a relational database tool, designed to capture and assimilate the metadata associated with a data warehouse project. Metadata allows the developer of the warehouse to understand the data and how it is used at the source, as well as how it is to be used in the warehouse. The metadata should be captured from documentary sources as well as individuals who know and understand the data. The knowledge that is captured can then be used to make decisions as to which data fields should be processed further to attain the quality necessary to serve the purposes of the warehouse.

Several unique challenges are encountered in assessing the metadata in a warehouse. Metadata in a data warehouse is much broader than the physical characteristics typically described in a data dictionary. The data dictionary "is a collection of descriptions of the data objects or items in a data model for the benefit of programmers and others who might need to refer to them (whatis.com 2000)." The data warehouse draws on multiple sources to create a repository designed for management decision support (Inmon, 1996). The metadata for a data warehouse must contain information that allows for integration of data across these multiple sources. Data from the multiple sources may exhibit semantic differences (Ballou and Tayi, 1999). For example, one source database may refer to the gender of an individual as Gender, another as Sex, and yet another as "ADHSEX ."[1] An analysis of the various data sources is needed to highlight fields where similar data are collected in fields with different names. Additionally, data fields with similar names may contain data that are quite different. For example, a field named "Contract" may contain the names of the people associated with the contract in one database, the status of a contract in another database, and the date of the contract in a third database. Yet another problem is the fact that even when two data fields have the same name, i.e. Gender, and collect the same data, i.e. male and female, it may be represented in the source databases differently, i.e. M/F, 0/1, or Male/Female. The DQKM is designed to address these semantic issues, commonly known as homonyms and synonyms.

The quality of the data, as noted earlier, is dependent on its use and the quality dimensions being considered. A measurement can be assigned based on the data field, the quality dimension, and the use of the data in the warehouse. Individuals who are responsible for the data at the source (typically those who interact with the data on a regular basis), in conjunction with the individuals who understand how the data will be used in the warehouse (typically the warehouse development team), will be responsible for determining this subjective measurement. The measurement in the data warehouse may or may not be what the measurement is in the source database. For example, the Ethnicity field in the source may be only 50% accurate, based on the fact that only 50% of the data in the records contains the detailed information that is required at the source. However, for determining client demographics (the use), if the ethnicity is "in the ballpark", it is good enough. For the purposes of client demographics, the various sub-categories of Hispanic will be aggregated, whereas a specific country of origin (Cuban or Puerto Rican, for example) had been required at the source. Thus, for this use (client demographics) and the accuracy dimension, the measurement may be 90%. In other words, 90% of the data in the source is accurate when rolled up into the aggregate.

The relational schema for the DQKM is presented in Figure 1 on the next page. Two entities are highlighted for further clarification, the DATA Entity and the IMPORTANCE Entity.

DATA Entity

The DATA Entity focuses on the physical characteristics of the data, collecting much of the same information that would be found in a standard data dictionary. However, three of the fields address the unique needs of a data warehouse. For instance, the **Skip?** field would be used to indicate that a specific data field should not be migrated to the warehouse. This determination may be made before or after the measurement is assigned. If the reason for a "yes" in this field is that the data does not address the needs of the warehouse, no measurement need be assigned. On the other hand, if the measurement does not meet the minimum criteria for further enhancement, then this box would be checked at a later point in the process.

Data_Category is a drop down list that is customized to the data warehouse and source databases for the project. The data category field for the HIMS project (the case study used to populate the DQKM) contains records that identify client demographics, shelter characteristics and data necessary to calculate length of stay. Specific data categories included gender, ethnicity, admit date and discharge date, among others. The purpose of the data category field is to highlight synonyms. In populating the DQKM, the user will code fields with the appropriate data category. Thus, if several data sources contain data

[1] "ADHSEX" is a field name from a source database used to populate the Data Quality Knowledge Management (DQKM) tool. It is used here as an example to show the ambiguity of many field names from legacy systems.

related to gender, but have different names (sex, gender, etc.) then the **Data_Category** field can been used to perform SQL queries and obtain all available fields for gender, regardless of what they are called. The decision as to which category a given source data field corresponds to is a decision made by the individual responsible for populating the DQKM.

Finally, **Field_Type** carries a drop down box of 7 data types, similar to the types used by Foley in the Ardent Quality Manager methodology (Foley, 1999). These field types include binary, coded, date, formatted text, free text, key and numeric. The field type can help the analyst determine whether the data should be cleansed manually or automatically. In order to clean the data automatically, there must be something to compare to. If it is a coded field, a key field, a date or numeric field, then it can potentially be cleaned using a data quality tool.

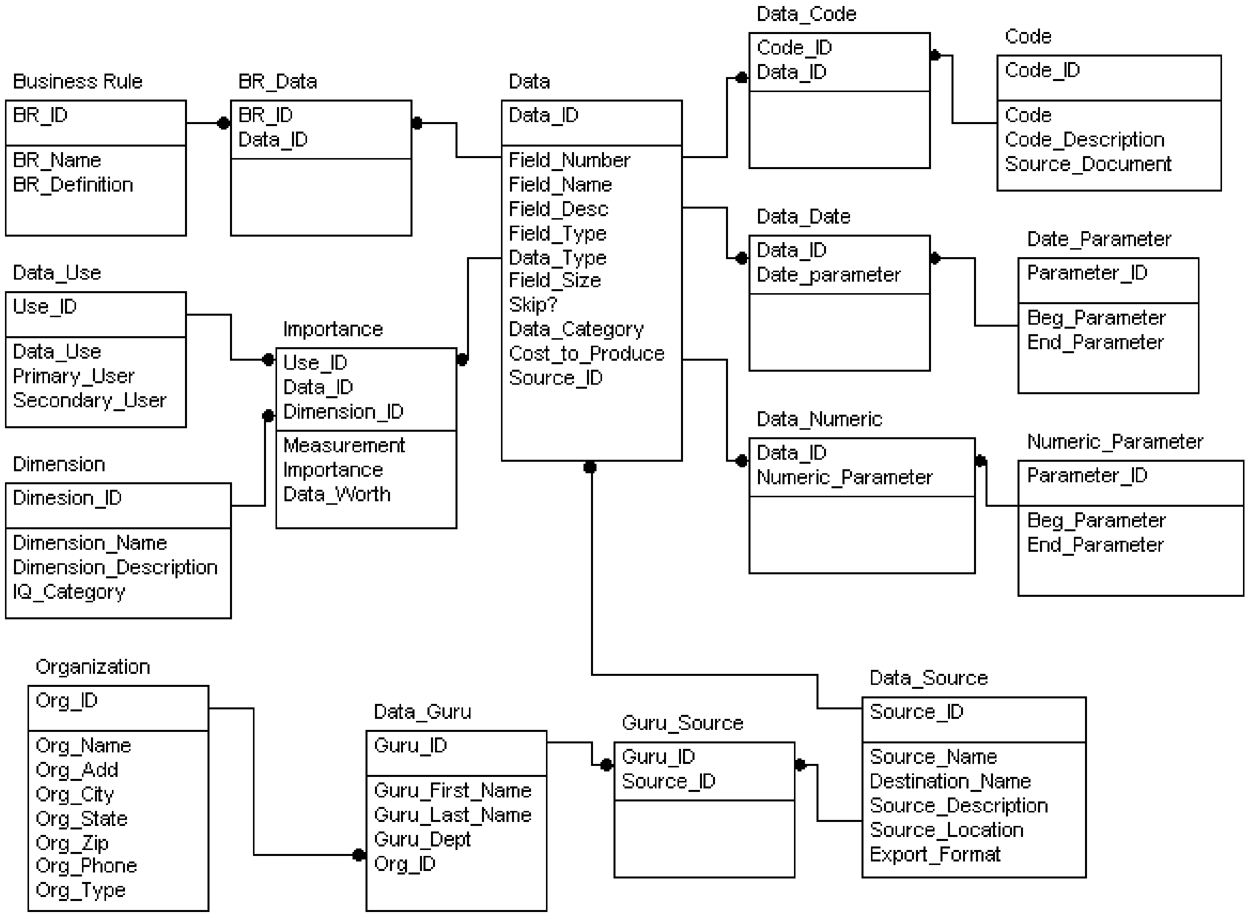


Figure 1. Data Quality Knowledge Management (Dqkm) Schema (Neely, 2001)

IMPORTANCE Entity

The IMPORTANCE table is in a ternary relationship with USE, DIMENSION and DATA. This entity takes into account the fact that you must know both the use and dimension attributes of the data in order to determine its importance in the data warehouse. It captures the "fitness for use" characteristic (Tayi and Ballou, 1998). This field will have a measurement, dependent on its use and dimension. Typically this is a subjective measurement. For example, for a given data field, the completeness dimension for client demographics may be 50%, whereas the accuracy dimension for client demographics may be 85%. The example used previously for client demographics is an example of a subjective measure. Occasionally, the measure may be objective depending on the use and dimension. For instance, on the "Specific Shelter ID" use, it would be possible to calculate the number of empty fields in a source database to arrive at a completeness measure. The non-null fields would then be a percentage of the total. The measurement in this table may or may not be what the measurement is in the source database.

The DQKM will be tailored to meet the needs of the specific warehouse project that is being conducted. For the HIMS prototype the DQKM included data uses as seen in Table 1. Data Uses should specifically address the needs of the warehouse being designed. The prototype was intended to answer the general question of whether services provided to homeless individuals were making a difference in the recidivism rate. Thus client demographics, services and length of stay uses, among others, were created.

Table 1. Data Use Drop-Down Box for DQKM

Use-ID	Data_Use	Data_Use_Description	Primary_User
1	Client Demographics	Attributes of the population	Caseworker
2	Services Assessed	Determination of services assessed to an individual client	Caseworker
3	Services Offered	Determination of services to be offered by the sate	State and agencies providing services
4	Shelter Characteristics	Identification of shelter characteristics (single vs. family, male, vs. female, etc.)	Shelter owners
5	Length of Stay	Determining lenth of stay	THA, shelter owners
6	Specific Shelter ID	Identification of specific shelters in the system for anlaysis	State

The Dimension table is drawn from the work of Wang and Strong (1996). As seen in Table 2, there are 15 dimensions of data quality. As noted earlier, there will be trade-offs in these dimensions. Thus, a given data field will be evaluated for quality on one or more specific dimensions. Separate measurements will be assigned for each combination of dimension and use. Thus, a measurement may be assigned for gender on the accuracy dimension and the completeness dimension.

Table 2. Dimension Table of the DQKM (Wang and Strong, 1996)

Dimension _ID	Dimension_Name	Dimension_Description
1	Believability	believable
2	Accuracy	data are certified error-free, accurate, correct, flawless, reliable, errors can be easily identified, the integrity of the data, precise
3	Objectivity	unbiased, objective
4	Completeness	bredth, depth and scope of information contained in the data
5	Reputation	reputation of the data source, reputation of the data
6	Value-added	data give you a competitive edge, data add value to your operations
7	Relevancy	applicable, relevant, interesting, usable
8	Timeliness	age of data
9	Amount of Data	appropriate amount of data
10	Interpretability	interpretable
11	Ease of Understanding	easily understood, clear, readable
12	Representational Consistency	data are continuously presented in same format, consistently represented, consistenlty formatted, data are compatible with previous data
13	Concise Representation	well presented, concise, compactly presented, well-organized, aesthetically pleasing form of presentation, well formatted, format of the data
14	Accessibility	accessible, retrievable, speed of access, available, up-to-date
15	Access Security	data cannot be accessed by competitors, data are of a proprietary nature

To reiterate, the measurement is assigned to a given data field based on its use and dimension. This measure is frequently subjective, but occasionally objective. The measure is a cooperative assignment, utilizing the expertise of the data providers and the data warehouse developers.

Exploratory Study Motivation

The DQKM was populated with metadata drawn from a case study that was described in a tutorial at AMCIS 1999 (Neely and Pardo, 1999). In the case study, the development of a prototype data warehouse, the Homeless Information Management Systems (HIMS) is described. The metadata that was used to design and create the HIMS prototype was used as the dataset for the exploratory study phase of this research. Source documents relating to homeless shelter providers and the Bureau of Shelter Services (BSS) were used to populate the DQKM.

Several characteristics of this case study make it a solid choice for this purpose. First, this case is a good candidate for the study due to the fact that the agency that was initially interested in the project had only a small piece of the puzzle needed to answer the questions regarding recidivism. The homeless shelters and other agencies providing services had the other necessary pieces. Thus, source databases for the warehouse came from a variety of organizations. This is a good example of the issues that are faced in developing an integrated repository. As a result, source documentation was available from six source databases. This documentation consists of word processing files, codebooks, partial data dictionaries, and Excel spreadsheets. The multiple databases allow the DQKM to be populated with metadata that will allow the user to evaluate homonyms and synonyms.

The variety of data in the HIMS prototype also caused this to be a good choice for testing the DQKM. Many of the warehouses that the interviewed experts were developing consisted primarily of financial data. The HIMS prototype contains demographic data, as well as financial data. The measurements assigned to data fields, based on data use and dimension, are frequently the same in the warehouse as they are at the source for financial data. This is because a key concern for financial data is accuracy and completeness, both of which need to be near 100% at both the source and the target. On the other hand, demographic data can be evaluated on many dimensions, and the quality at the source does not necessarily need to be the same as the quality in the warehouse. Finally, this researcher was very familiar with the HIMS metadata. This characteristic allowed for population of the DQKM in a manner that was analogous to population in a real-world data warehouse project.

Exploratory Study Process

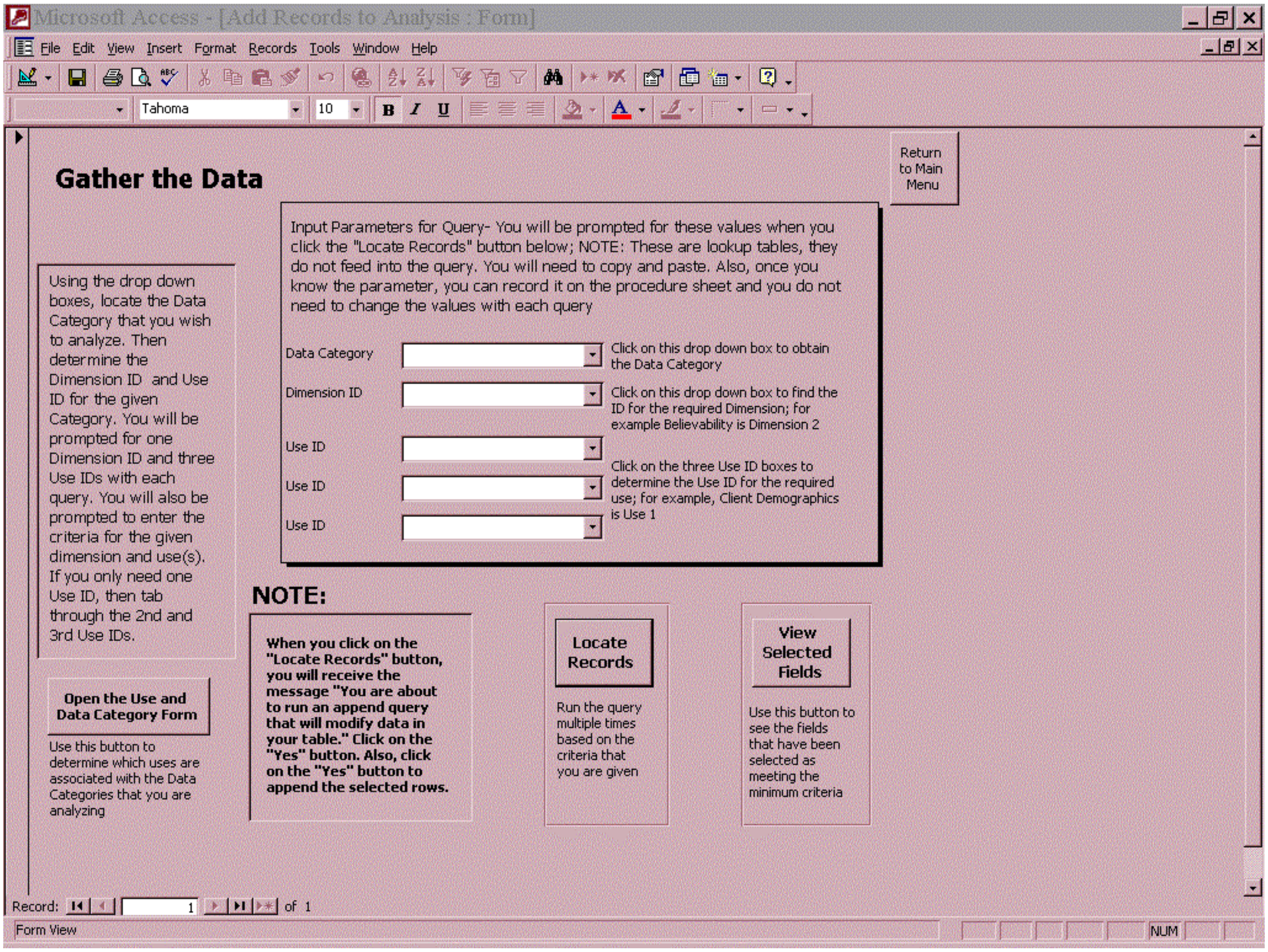
As indicated, an exploratory study was conducted, using the populated DQKM, in an effort to determine if the DQKM could be used to make resource allocation decisions regarding data quality projects. The subjects used in the study were senior students in a special topics class on Data Quality at Marist College. The focus of the study was on the IMPORTANCE Entity. On the assumption that data of a given (poor) quality is not cost effective to enhance, subjects were provided with a set of criteria on which a given data field, for a given use and dimension, were to be evaluated. (See Table 3 for the criteria used in the study.) If the specific data field met the minimum criteria, it would be evaluated further for quality enhancement. Students were instructed to use the DQKM to extract specific data fields that met the minimum criteria. See Figure 2 on the next page for a screen shot of the data entry form the students used to capture these fields.

The output from student's efforts was extracted to a new table, which could then be further analyzed. The students were asked to export this table to an Excel worksheet and sort the data based on the following priorities for sorting the data fields that met the minimum criteria for further enhancement:

- If a data field (identified by the Data_ID label) has multiple uses it has the first priority.
- Data used for **length of stay** (one of the uses) has a higher priority than the other uses
- Data that is closer to 100% in a given dimension will require fewer resources and thus, should be given higher priority

Table 3. Criteria for Further Consideration of Quality Enhancement

Dimension	Use	Criteria
Complete	Client Demographics, Services Assessed, Services Offered	50%
Complete	Shelter Characteristics, Length of Stay, Specific Shelter ID	60%
Accurate	Client Demographics	75%
Accurate	Services Assessed, Services Offered	50%
Accurate	Shelter Characteristics, Length of Stay, Specific Shelter ID	60%
Reputable	Client Demographics	40%
Reputable	Services Assessed, Services Offered, Shelter Characteristics	55%
Reputable	Specific Shelter ID	80%
Accessible	All Uses	50%


Microsoft Access - [Add Records to Analysis : Form]
File Edit View Insert Format Records Tools Window Help
Tahoma
10
B I U
Return to Main Menu
Open the Use and Data Category Form
Locate Records
View Selected Fields
Record: 1 of 1
Form View
NUM

Gather the Data

Using the drop down boxes, locate the Data Category that you wish to analyze. Then determine the Dimension ID and Use ID for the given Category. You will be prompted for one Dimension ID and three Use IDs with each query. You will also be prompted to enter the criteria for the given dimension and use(s). If you only need one Use ID, then tab through the 2nd and 3rd Use IDs.

Input Parameters for Query- You will be prompted for these values when you click the "Locate Records" button below; NOTE: These are lookup tables, they do not feed into the query. You will need to copy and paste. Also, once you know the parameter, you can record it on the procedure sheet and you do not need to change the values with each query

Data Category — Click on this drop down box to obtain the Data Category

Dimension ID — Click on this drop down box to find the ID for the required Dimension; for example Believability is Dimension 2

Use ID

Use ID

Use ID

Click on the three Use ID boxes to determine the Use ID for the required use; for example, Client Demographics is Use 1

NOTE:

When you click on the "Locate Records" button, you will receive the message "You are about to run an append query that will modify data in your table." Click on the "Yes" button. Also, click on the "Yes" button to append the selected rows.

Use this button to determine which uses are associated with the Data Categories that you are analyzing

Run the query multiple times based on the criteria that you are given

Use this button to see the fields that have been selected as meeting the minimum criteria

Figure 2. Gather the Data Input Form for the DQKM

Results

The results of the study indicated that the DQKM could be used successfully to extract the data fields that met minimum criteria. However, the subjects were unable to prioritize the data fields to allocate the $50,000 budget for further enhancement. Students were not equipped to complete the prioritization phase for a variety of reasons. The factors that played a role in this study included: lack of analysis experience, lack of training on the DQKM and Excel, and difficulty with the user interface. Future studies with the tool will control for these factors to determine if one factor was more significant than another.

A decision had been made prior to this study that the students would perform the tasks individually. With a class size of 15, it was hoped that individual results would provide more interesting data. However, the results indicated that it may have been more advantageous to have students working in groups, with the ability to interact with each other, as well as the manager. Future research with this tool will use a larger subject population. However, for exploratory purposes, the findings were interesting and warrant further investigation.

Conclusion

As described in this paper, a relational database tool, the Data Quality Knowledge Management (DQKM) captures and organizes the metadata associated with a data warehouse project. It builds on the concept of fitness for use by describing a measurement technique for subjectively assigning a measure to a data field based on the use and quality dimension of the data within the data warehouse. This measurement can then be compared to some minimum criteria, below which it is not cost effective to enhance the quality of the data. This tool can be used to make resource allocation decisions and get the greatest benefit for the cost in utilizing the scarce resources available to enhance source data for a data warehouse.

References

Ballou, D.P. and Tayi, G.K. "Methodology for Allocating Resources for Data Quality Enhancement," *Communications of the ACM* (32:3), 1989, pp. 320-329.

Ballou, D.P. and Tayi, G.K. "Enhancing data quality in data warehouse environments," *Communications of the ACM* (42:1), 1999, pp. 73-80.

Foley, E.A. "Ardent Data Quality Analysis Methodology," Ardent Software, Inc, 1999.

Inmon, W.H. "The Data Warehouse and Data Mining," *Communications of the ACM* (39:11), 1996, pp. 49-50.

Iverson, D. "Meta-Information Quality, Keynote Speech, IQ2001," Cambridge, MA,), 2001,

Neely, M.P. "A Proposed Framework for the Analysis of Source Data in a Data Warehouse," *Proceedings of the Proceedings of The Conference on Information Quality*, MIT, Cambridge, MA, 2001,

Neely, M.P. and Pardo, T.A. "Teaching Data Quality Concepts Through Case Studies," *Proceedings of the Proceedings of Americas Conference on Information Systems*, Milwaukee, WI, 1999,

Redman, T.C. "The Impact of Poor Data Quality on the Typical Enterprise," *Communications of the ACM* (41:2), 1998, pp. 79-82.

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

Tayi, G.K. and Ballou, D.P. "Examining Data Quality," *Communications of the ACM* (41:2), 1998, pp. 54-57.

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

Wixom, B.H. and Watson, H.J. "An Empirical Investigation of the Factors Affecting Data Warehousing Success," *MIS Quarterly* (25:1), 2001, pp. 17-38.