A DATA QUALITY ENGINEERING PROCESS

Donna M. Meyen
The MITRE Corporation

Mary Jane Willshire, Ph. D.
Colorado Technical University

ABSTRACT

In support of the United States Air Force, The MITRE Corporation developed a top-level process and method for understanding, managing, and directing the architectures, interoperability, and evolution of complex systems of systems. Data that were collected on these systems over a two year period are now viewed with concern as to their quality. For example, the accuracy and timeliness of the data are suspect because the data have not been updated since their initial collection. Out of this the USAF introduced a requirement for quality data. Research into data quality and data quality engineering arenas has revealed that there is no strategic solution to the entire problem; a data quality process does not exist. The purpose of this research is to present a Data Quality Engineering Process (DQEP) into which the various single solutions can be integrated in order to effect data quality process and product improvement.

A Data Quality Engineering Process

1. Introduction

Over the past two years, The MITRE Corporation has been collecting extensive data on its project to support the United States Air Force (USAF) with a top-level process and method for understanding, managing, and directing the architectures, interoperability, and evolution of complex systems of systems. Recently, various users of the data have raised concerns regarding the quality of the collected data. For example, the accuracy and timeliness of the data were suspect because the data have not been updated since their initial collection. Even though the USAF recognizes the importance of these data to strategic planning, shrinking resources forced the USAF to focus available resources on activities other than the maintenance of these data. However, the USAF's requirement for strategic planning remains; therefore USAF's dependence on these data remains. Thus, the USAF's requirement for quality data emerged.

Research into the identification of candidate data quality engineering processes revealed that various facets of data quality have been investigated and reported on in numerous fields such as chemistry [17], education [19], accounting [4], petroleum engineering [9], and law enforcement [18]. Mathematicians, information management experts, and computer scientists have offered numerous ways in which to define [8, 14], analyze [15, 16] and improve the quality of data [1, 2], regardless of the format, media, or content. Both automated tools [3] and formal methods for data comparison [12] are available. But none of the mentioned approaches organizes these partial solutions in a coherent fashion to provide an overarching framework in attempts to permanently solve a data quality problem. In short, no one has attempted to develop an entire data quality engineering process or program from beginning to end.

This paper is a report on research in progress that attempts to develop a general Data Quality Engineering Process (DQEP) that can be used to define, analyze, and provide guidance to improve data quality. Work completed as of August 30 is described here and further advances will be discussed during the presentation. The paper is organized as follows: Section 2 discusses the general method for developing and applying the DQEP; Section 3 provides a brief case study in one domain; and Section 4 concludes the paper.

2. The General Method

Our method to develop and prove the viability of the DQEP consists of six steps, each of which is described in more detail in this section. Not all steps are complete at the time of this writing. In this paper we describe the completed portions of our work and outline the remaining effort.

Using a wide range of sources expert in the area of data quality and best engineering judgment, the DQEP has been developed so that data quality is treated consistently within the context of the specific systems and their functions, the data at hand, and most importantly, the users of the data. There are four major steps in the DQEP: define a Data Quality Engineering Model (DQEM); define the relevant data quality attributes; analyze the data quality attributes; and provide data quality improvement guidance. Of these four steps, we have defined the DQEM as discussed below.

The first step in DQEP is to develop an appropriate functional and data modeling paradigm to produce a DQEM. The domain that is supported must be described in detail, so the context within which the data operate is clearly understood. The primary and all non-primary users of the data must also be described in the DQEM. Once the scope of the domain is established, all other tasks are accomplished within this very specific context. This includes defining information and data requirements, defining information and data flow, and defining business rules. Next is the precise definition of the data quality parameters peculiar to the application and the value of the information. Data quality parameters are selected and defined based on the users' requirements, functional context, recommendations from the literature, and other expert advice. Then the methods of collection and measuring data quality are selected and initial data quality measurements, using at least two different methods, are recorded. An interpretation task is then initiated to identify problem areas and their probable and possible causes. Finally, we must identify and select remedies to incorporate into the data life cycle. After some predetermined interval or event, a second set of measurements is taken using the same methods as for the initial measurements. An analysis of the impact of the adjusted data life cycle on the quality of data is undertaken to reveal any areas requiring additional attention. The results of this analysis are used to refine and adjust the model.

The second step in developing and proving the DQEP is to define measures of success. These are the criteria used to judge whether or not the DQEP is successful. These include such things as:

1. The general process was successfully used as a basis to define data quality attributes specific to a particular domain and users' needs.
2. The general process was successfully tailored to facilitate the measure and analysis of the quality of a specific set or sets of data.
3. The general process provided guidance in identifying methods to improve the quality of data in a selected case or cases.

Qualitative and/or statistical evidence must be presented to substantiate the findings. Measurable thresholds must be established against which all results are compared. Qualitative measurements may also be used within the specific context of the data function and usage.

The third step is to execute the DQEP using new or selected portions of existing databases as test cases in order to verify the DQEP. The functions requiring the services of the associated databases are described, and the services provided by the databases to these functional areas are also documented. The logical structure of the databases and the appropriateness of applying the DQEP to those databases are discussed. This provides the context for tailoring the DQEP based on the descriptions in the DQEM.

Available alternatives for each DQEP task are evaluated in light of the domain and the rationale for selections is documented.

In step four, the results of applying each of the DQEP's activities are presented and analyzed in detail in order to validate the process. This demonstrates that the DQEP is a viable framework for our task. Comparison of data quality measurements is presented here, along with an analysis of the root causes of any discovered problems. Attempts at remedies are documented here, along with the results of their implementation. At this point, any problems with the DQEP and/or data and the data life cycle indicates that the definition of the DQEP may require adjustment. This may take the form of either relaxing or tightening the data quality requirements.

Conclusions are drawn in step five of the method. A determination is made as to whether the success criteria for developing the DQEP have been met. If, using quantitative and qualitative measurements, it is determined that the criteria have been met, then it follows that a viable DQEP has been successfully developed and applied. Other conclusions that may be made at this point include any general limitations inherent in the DQEP, the cost/benefit ratio of the implementation of a data quality improvement program, and constraints imposed on the model by staff considerations such as skill level, availability, etc.

The sixth and final step in the method is to offer recommendations indicated as a result of the previous five steps. Areas where recommendations may be offered include:

1. The specification of requirements necessary to implement the DQEP – process engineering, staffing, skill levels, automated tools, time cost, and any other significant or extraordinary resource requirements discovered during the execution of the DQEP are documented in this section.
2. Technology insertion, including both hardware and software – the identification of areas in the DQEP where technology insertion would improve the process (and therefore the product) is provided here.
3. Areas warranting further investigation – once all six major steps of the method to develop and apply the DQEP have been successfully executed, many of the theories offered in the current literature will have been applied in a real-world setting. The results of implementing these theories in a practical application provide valuable insights as to the focus of future research and application of data quality engineering methods. These insights are documented in this section.

The remainder of the paper provides a brief case study where this method has been applied in the context of the Air Force (AF) Command, Control, Communications, Computers and Intelligence (C^4I) domain.

3. A Specific Application of the Method

This section briefly describes the tailored DQEP for the AF C^4I domain. The context and methods used in the six steps involved in developing, applying, and proving the viability of the DQEP are discussed. The modeling element of the DQEP for AF C^4I interoperability domain has been completed. The remaining three elements of the process (define, analyze, and improve) will be tailored to this domain in the near future and will be documented as the project continues.

3.1 Developing the DQEM for the AF C^4I Database

3.1.1 Domain Description

The mission and support areas of the USAF include Combat Operations, Intelligence Support, Mission Support, Mobility Operations, Space Operations, and Special Operations [26]. For these areas, the Air Force is the central focus, but it operates within the boundaries of the Department of Defense (DOD), which in turn operates within the boundaries of the non-DOD world. Each of these mission and support areas is supported by Command and Control (C^2) systems consisting of interacting elements. A model of the types of elements (or nodes) and the interactions that occur between them are shown in Figures 1 and 2. There are four types of nodes: C^2 System, Data Source, Resource/Asset, and External Agency.

A C^2 System is defined as a node where activities such as data fusion, information processing, and dissemination are performed. These functions support the major function of a C^2 system, that of decision making. Data Sources provide "root and cause" information and data to the system, such as warning information from sensors, weather data, or satellite photographs. A Resource/Asset is a center that executes an order or directive received from a C^2 element. These include directives to ship supplies to specified areas, apply weapons, or change the position of an orbiting satellite. An External Agency is an organization whose primary mission is not C^2, but which may influence the decisions made by the C^2 systems (e.g., Congress).

The description of any two nodes plus the communications between them is called an intersection and is described by the sub-model depicted in Figure 2. Attributes include the name of the source node, the name of the destination node, the mission accomplished by the passing of data, the Automated Information System (AIS), if any, and the communication network over which the data are passed. The most important nodes in performing each of the six Air Force missions have been identified, as have their intersections with other nodes. These intersections occur within a specific mission area (intra-mission), between AF mission areas (inter-mission), or between AF and non-AF areas (extra-mission). All of the intersection information for the top level and expanded model has been collected and resides in a relational database, the AF C^4I Database. Figure 3 depicts the structure of the core table maintained by this database called the Interoperability Table. For each intersection, the table contains the attributes as depicted in the figure. To illustrate, Table 1 contains some of the definitions of data elements used in the Interoperability Table.

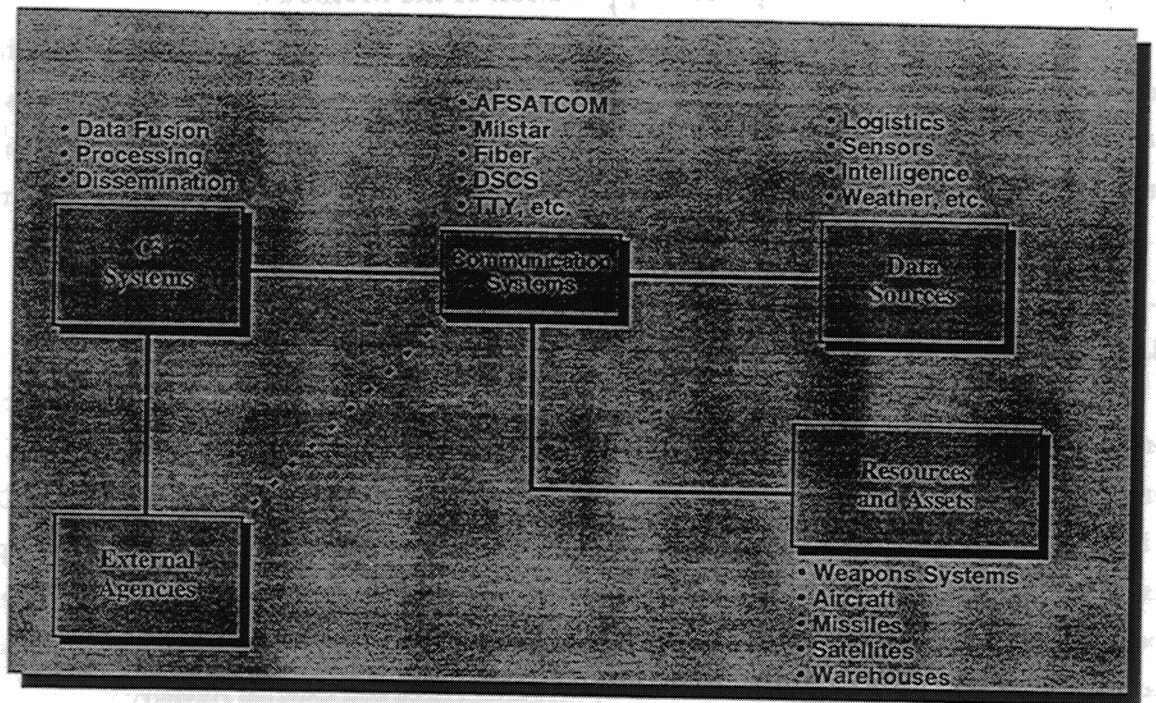

Figure 1. The Air Force C[4]I Model

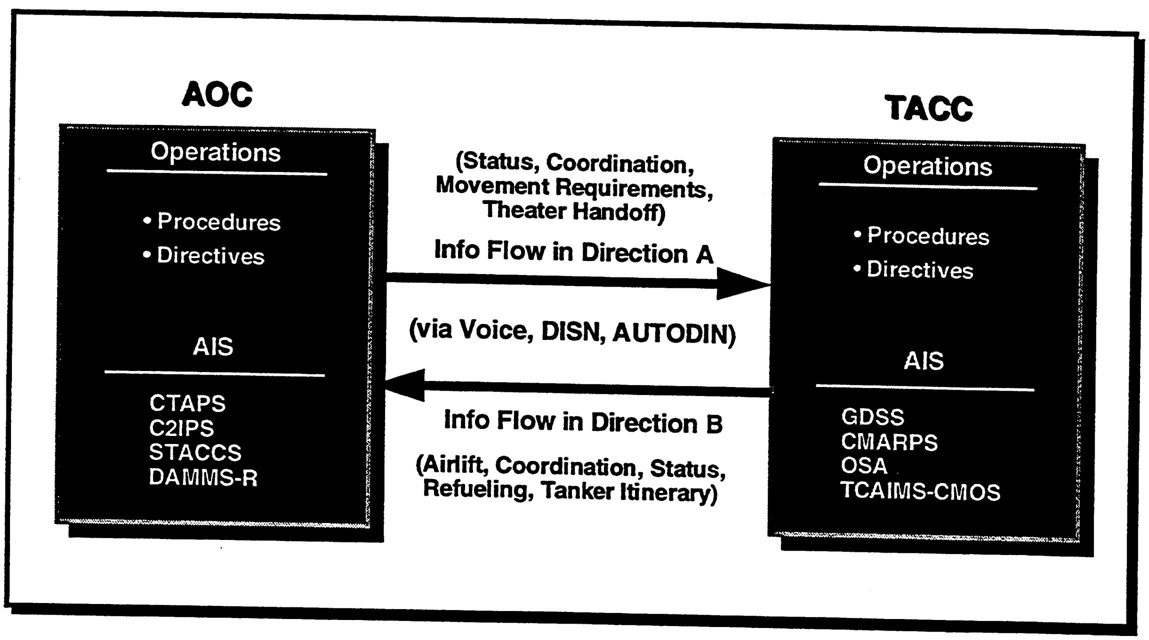

Figure 2. An Intersection with Attributes

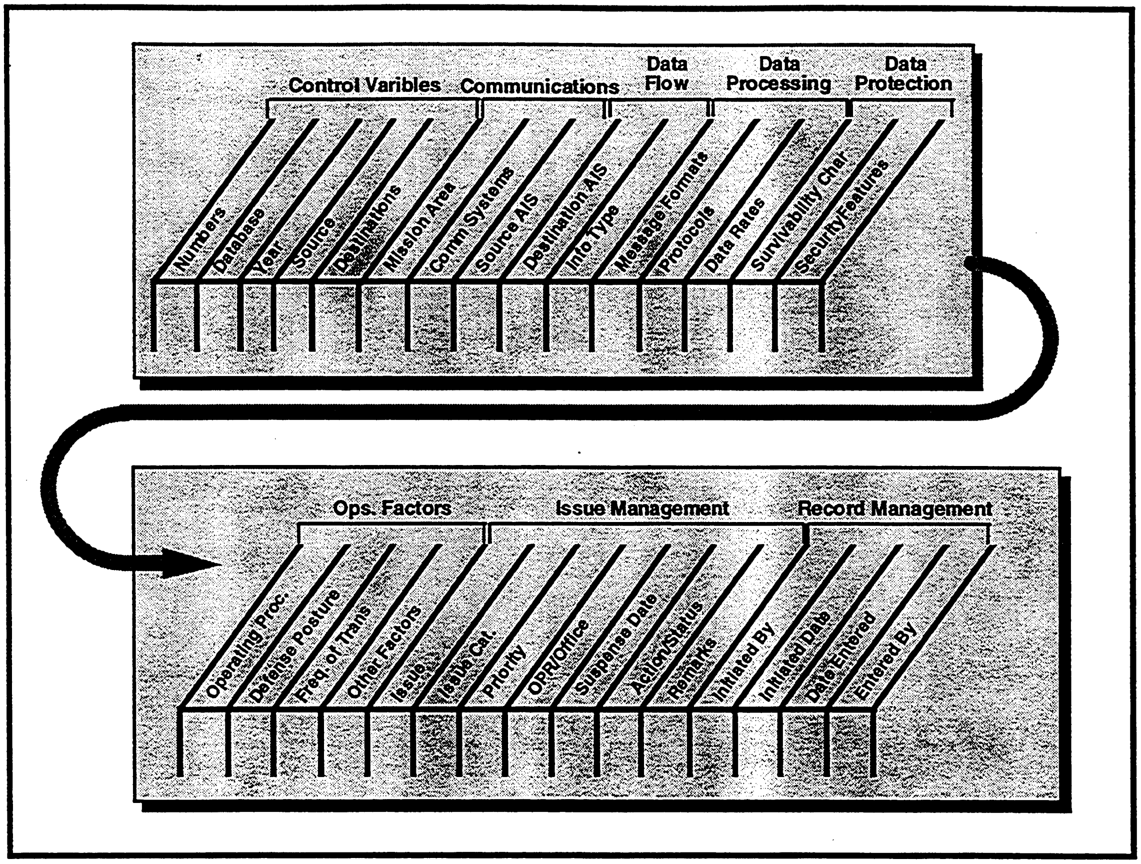

Figure 3. Interoperability Table Content

We continue building our model by considering the users of the data. The primary users of the AF C^4I data are the USAF Directorate of Communications and Information (USAF/SCT) and their technical arm, the Air Force Communications Agency (AFCA). The major mission of these two organizations is to determine inter- and extra-mission interoperability requirements, issues, and resolutions. The secondary users of the AF C^4I data are Air Mobility Command (AMC), Air Force Special Operations Command (AFSOC), Air Force Space Command (AFSPC), Air Combat Command (ACC), Air Force Materiel Command (AFMC), United States Air Forces Europe (USAFE), and Pacific Air Force (PACAF). These command- and theater-level users employ the AF C^4I Database to determine intra-mission interoperability requirements, issues, and resolution. In addition, the AF C^4I Database is used to identify system-level interfaces to other mission area systems. Tertiary users of the AF C^4I data are organizations such as Electronic Systems Command (ESC), the major system acquisition agency of the Air Force.

Figure 4 depicts the data life cycle for the AF C^4I Database. There exists a Master AF C^4I Database where all AF mission data reside. Horizon Link (HL) is a relational database manager specifically developed for this application that provides automated manipulations of the AF C^4I Database. Of particular interest to this research is the AF C^4I Database maintenance process; this maintenance process forms the baseline of the DQEM under investigation.

Abbreviation	Name	Description	Type[1]
COMM_NTWK	Communications Network	List of the physical communications systems used in the transmission of data between the elements.	254 A/N
DESTIN	Destination	The end point or element of data receipt.	60 A/N
DEST_AIS	Destination Automated Information System (AIS)	List of major AISs available to the destination element to receive and process information.	254 A/N
DEST_ID	Destination Identification (ID)	ID of destination element.	6 N
DIAGRAM	Diagram	Diagram for which the intersection is applicable. Values: 1 = Top Level 2 = Combat Operations 3 = Intelligence Support 4 = Mission Support 5 = Mobility Operations 6 = Space Operations 7 = Special Operations 8 = Korea 9 = PACAF 10 = USAFE	2 N
ENTEROR	Enteror	Logon identification of person who entered the intersection data into the database.	70 A/N
ENT_DATE	Date Entered	Date the intersection information was entered in the format MM/DD/YY, e.g., 12/31/94	Date
ENT_TIME	Time Entered	Time the intersection information was entered.	8 A/N
INFO_TYPE	Information Type	Brief description of the information passed between the source and destination elements.	254 A/N
INT_ID	Intersection Identification	Unique identification of intersection in form: NNNNA. The last character of the identification indicates direction (i.e., 0000a indicates the information is from the first element to the second element; 0000b indicates the information is from the second element to the first element).	5 A/N
POC	Point of Contact	Person or agency responsible for intersection	70 A/N
SOURCE	Source	The point or element of data origination.	60 A/N
SOURCE_AIS	Source Automated Information System	Major AISs available to the source element to process and transmit information.	254 A/N
SOURCE_ID	Source ID	Identification of source element.	6 N

Table 1. Interoperability Information in the Database

[1] A/N - Alphanumeric
N - Numeric

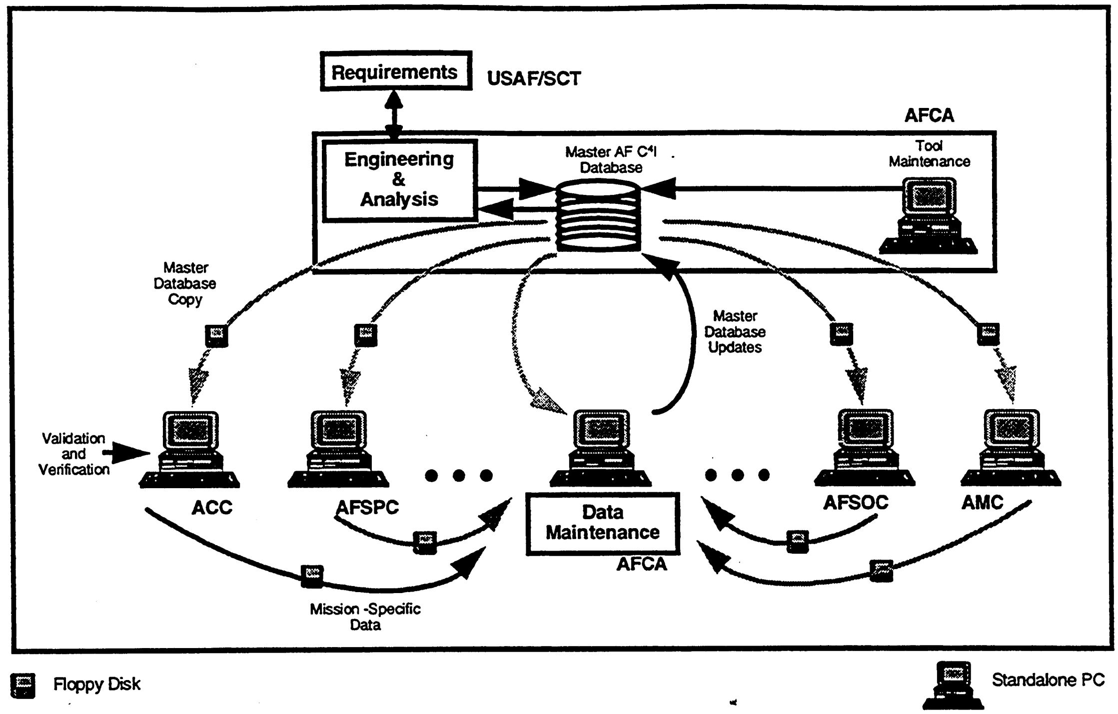

Figure 4. Air Force C[4]I Data Management Concept of Operations

The AF C[4]I Database maintenance begins when AFCA distributes copies of the master database on floppy disks to the individual Major Commands (MAJCOMs) and agencies noted previously. A run-time version of HL accompanies the database. This is depicted by the light gray arrows on Figure 4. Each Office of Primary Responsibilitiy (OPR) loads HL and the source data on a standalone PC and modifies the data for which that organization is responsible. The OPR is also responsible for the validation and verification of its respective information. Changes are recorded by HL, the appropriate tables and graphics are copied to floppy disks, and then forwarded to a Central Maintenance Organization (CMO), namely AFCA. The CMO then merges the data gathered from the MAJCOMs into a new version of the master database. The cycle is then repeated. This process is considered to be the baseline DQEM.

3.1.2 Definition of the Data Quality Parameters for the AF C[4]I Database

Now that we have established our baseline DQEM, we must define the data quality parameters which are peculiar to this application, and the users of the data. This work is in progress and what follows is an outline of how we will proceed for this step. To do so we will select data quality parameters based on the users' requirements such as data criticality, functional context, and expert opinions.

In an effort to standardize the vocabulary of data quality, attempts have been made to define generic data quality attributes [10, 24, 27]. In a recent survey, a list of 179 data quality attributes has been compiled and categorized [23]. This comprehensive list, presented in Figure 5, will provide the basis for the selection of data quality attributes which are pertinent to the users described previously. Those items marked with an asterisk are under consideration for this project. As the final selections are made, precise

Ability to be Joined With	Ability to Download	Ability to Identify Errors*	Ability to Upload
Acceptability	Access by Competition	Accessibility*	Accuracy*
Adaptability	Adequate Detail*	Adequate Volume	Aestheticism
Age*	Aggregatability*	Alterability*	Amount of Data*
Auditable*	Authority*	Availability*	Believability
Breadth of Data*	Brevity	Certified Data*	Clarity*
Clarity of Origin*	Clear Data Responsibility*	Compactness	Compatibility
Competitive Edge	Completeness*	Comprehensiveness*	Compressibility
Concise*	Conciseness*	Confidentiality	Conformity*
Consistency*	Content*	Context*	Continuity
Convenience	Correctness*	Corruption*	Cost
Cost of Accuracy	Cost of Collection	Creativity	Critical*
Current*	Customizability	Data Hierarchy	Data Improves Efficiency
Data Overload*	Definability*	Dependability*	Depth of Data*
Detail*	Detailed Source*	Dispersed	Distinguishable Updated Files
Dynamic	Ease of Access*	Ease of Comparison*	Ease of Correlation*
Ease of Data Exchange*	Ease of Maintenance*	Ease of Retrieval*	Ease of Understanding*
Ease of Update*	Ease of Use*	Easy to Change*	Easy to Question*
Efficiency	Endurance	Enlightening*	Ergonomic
Error-Free*	Expandability*	Expense	Extendibility*
Extensibility*	Extent	Finalization	Flawlessness*
Flexibility*	Form of Presentation	Format*	Integrity*
Friendliness	Generality*	Habit	Historical Compatibility*
Importance*	Inconsistencies*	Integration*	Integrity*
Interactive	Interesting	Level of Abstraction*	Level of Standardization*
Localized	Logically Connected*	Manageability*	Manipulable*
Measurable*	Medium*	Meets Requirements*	Minimality*
Modularity*	Narrowly Defined*	No lost information*	Normality*
Novelty	Objectivity*	Optimality	Orderliness*
Origin*	Parismony	Partitionability	Past Experience
Pedigree*	Personalized	Pertinent*	Portability*
Preciseness*	Precision*	Proprietary Nature	Purpose*
Quantity	Rationality	Redundancy*	Regularity of Format*
Relevance*	Reliability*	Repetitive*	Reproducibility
Reputation*	Resolution of Graphics	Responsibility*	Retrievability*
Revealing	Reviewability	Rigidity	Robustness
Scope of Info	Secrecy	Security	Self-Correcting
Semantic Interpretation	Semantics*	Size	Source*
Specificity*	Speed	Stability*	Storage
Synchronization*	Time-independence*	Timeliness*	Traceable*
Translatable	Transportability*	Unambiguity*	Unbiased
Understandable*	Uniqueness	Unorganized	Up-to-Date*
Usable*	Usefulness*	User Friendly	Valid*
Value*	Variability	Variety	Verifiable*
Volatility*	Well-Documented*	Well-Presented	

Figure 5. Generic Data Quality Attributes [23]

definitions for each data quality attribute will be recorded and a rationale for each choice documented. Both qualitative and quantitative attributes will be selected.

3.1.3 Collection, Measurement and Analysis of Data Quality Attributes

The next effort is to select the methods used to collect and measure data. This too is an ongoing effort. Up to four different methods will be used to record the initial data quality measurements. These methods will be chosen based on the suitability of the methods to the type of data under scrutiny and to user requirements. An interpretation task will then be initiated to identify problem areas and their probable or possible causes. The following paragraphs present the methods under consideration for this task.

The attributes selected by the process described above will be categorized as either external or internal as defined by Fenton in [6]. He contends that we "measure attributes of things." In this work he presents the notions of direct and indirect measurements, and also introduces the terms "internal attribute" and "external attribute". Internal attributes are directly measurable whereas external attributes can only be measured indirectly. We will analyze and record a classification of the attributes selected in a similar manner. In addition, a measurement scale as described by Fenton in [7] will be used as a framework within which attributes will be measured and compared. The scale types defined by Fenton that will be adopted for the DQEP are nominal, ordinal, interval, ratio, and absolute. Each data quality attribute selected for measurement will be judged to be either internal or external. If the attribute is considered external, then internal attributes that can be measured to indicate the extent of this specific external attribute will be identified. All internal attributes will then be assigned the type of scale to be used in their measurement. All data collection activities will be accomplished manually.

There are several general ways to measure the quality of data. One of these methods is a decision analysis approach described by Kaomea in [13]. This method computes the value of data qualities in a given decision scenario and assumes that the data quality attributes of interest are already known. This is not the situation for this project. However, at some point it is envisioned that the data contained in the AF C^4I Database will be considered in decisions about new system acquisitions, so this method is of interest.

Another method is extending the Entity-Relationship (ER) or the Object-Oriented (OO) model to include quality attributes [15, 25]. The original theory stated by Kon [15] defined a data quality parameter as a "qualitative or subjective dimension by which a user evaluates data quality." Kon extends the ER model as defined by Codd [5] to attach data quality parameters to the attributes of the entities defined in the relations. Data quality indicators are added and are assigned values. The user is then able to judge the quality of the data based upon the value of these parameters. Since the AF C^4I Database uses a relational data model, and already includes some data quality attributes, it seems that this method is appropriate for our use.

A third method is based on mathematical foundations as in Morey [21]. This work is based on the data life cycle in a transaction-based system. Morey defines three key measures: the transaction reject rate, the intrinsic transaction error rate, and the stored record error rate. He goes on to develop estimation formulas for error rates and how to apply them to an analysis of the effectiveness of error reduction mechanisms. Since the data life cycle in this work closely parallels that of the DQEM it is likely that a modified version of Morey's work will be integrated into our effort.

The final method is based on Ishii as presented in [12]. This method is qualitative and is referred to as a "reduction-based data quality calculus." Since this method focuses on qualitative attributes of data obtained from several different data sources, and AF C[4]I Database contains data from several sources, it is well suited to our needs. The method will allow us to arrive at an overall data quality value based on relationships among the data quality attributes as dictated by the users' context.

3.1.4 Analysis and Interpretation of Results

To complete this task, based on the initial data quality measurements, an analysis and interpretation task will then be initiated to identify problem areas and their probable/possible cause(s). The list of problem areas will be grouped by cause, which will then be prioritized. The objective will be to recommend modifications to the DQEM which will:

1. prevent further pollution of the database by eliminating/reducing the causes of low quality data;
2. provide each class of user with an indication of the quality of the data;
3. provide a means to identify and repair unacceptable values in the database, and to disseminate these values;
4. possess a optimal cost/benefit ratio.

We conclude this phase by identifying, evaluating and selecting remedies and incorporating them into the data life cycle. Drastic actions such as redefining the conceptual database design to include data quality requirements will be considered, as discussed in [22]. Re-design may be indicated in order to optimize those attributes deemed critical by all levels of users. Recommendations based on work by Ballou and Pazer [1], and McGee [20] will be considered. It may not be possible to determine all causes of errored data. In that case, a data tracking mechanism such as described in [11] may be considered. Lastly, a manual version of data scrubbing [3] may be necessary.

Once a set of remedies has been chosen, a plan will be devised and activated to incorporate those remedies. The plan will include the specific actions to be taken to incorporate the remedies, and will include the specific criteria which will define the successful implementation of the remedies. Once these criteria are met, a second measurement, using the same methods selected for the initial measurement, will be taken. An analysis of the impact of the adjusted data life cycle on the quality of the data will be done. The results of the analysis will be considered in the determination of the success or failure of the DQEP.

3.2 Success of the DQEP

The previous sections discussed the first major step in developing the DQEP. The next major step is to define a measure of success for the DQEP. There are three components we will address here. These success criteria are as follows.

1. The general process was successfully adapted to define data quality attributes specific to this domain and users. The classes of database users will be summarized here, along with the number and relevance of the data quality attributes specific to each class. A summary of the criticality of those attributes to each user class will be presented.
2. The general process was successfully tailored to facilitate the analysis and measurement of the quality of a specific set or sets of data. The different types of measurements will be listed here. A summary of the objective and subjective measurements will be tabulated. The number and types of major causes of data quality problems discovered during the process will be listed here as well.
3. The general process provided guidance in identifying a process to improve the quality of data in a selected case or cases. A comparison of the initial and second measurements will be included in this section. The definition for improvement in qualitative measures will be that the majority of those attributes showed an improvement of at least one level after remedial actions. For quantitative measures, improvement means that the majority of those attributes showed an improvement of at least 5% after remedial actions.

3.3 Verifying the DQEP

The third step is to execute the DQEP in order to verify it. Here the functions requiring the services of the associated databases will be described. The services provided by the databases to these functional areas will also be documented. The logical structures of the databases are given and the appropriateness of applying the DQEP to these databases will be discussed. Within this context, the DQEP will be tailored to the functions and data that are described in the tailored DQEM. The available alternatives for each DQEP task will be evaluated in light of the domain, users, and data; with the rationale for each selection documented. The exact steps executed, the means of accomplishment, and other external influences will also be described.

3.4 Analysis of the Results of Applying the DQEP

In step four, the results of applying the DQEP's activities will be presented and analyzed in detail. This provides the validation of the DQEP. Here we show that the DQEP is a viable tool to define, analyze, and improve data quality. A comparison of data quality measurements will be presented here along with an analysis of the root causes of any discovered problems. Attempts at remedies will be documented along with the results of their implementation. An analysis of the measures of success will reveal whether or not the DQEP was successfully developed. All three measures of success must be met for the research to be considered successful.

3.5 Conclusions for the DQEP

As step five in the method, conclusions will be drawn. First and foremost, a determination will be made as to whether or not the success criteria have been met. The quantitative and qualitative measurements will provide the foundation for drawing all conclusions. Other conclusions, besides the success of the DQEP, to be presented here may include any general limitations inherent in the process. This may include an inability to tailor for certain types of information bases. In addition there may be conclusions regarding the cost/benefit ratio for implementing this process and constraints imposed on the model by staffing considerations.

3.6 Recommendations for the DQEP

The final step of the method is to offer recommendations indicated as a result of executing the previous five steps. Areas where recommendations may be offered include: the specification of requirements necessary to implement the DQEP; technology insertion including both hardware and software; areas warranting further investigation.

4. Summary

This paper has described an approach to developing a general Data Quality Engineering Process (DQEP). The need for such a process was demonstrated, and the steps involved in developing, applying, and proving the DQEP as a viable process were described. A brief overview of an actual project which is in progress was given. For a more extensive discussion of the project and future results, please contact the authors.

5. References

[1] D. Ballou and H. Pazer, "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-95-01, January 1995.

[2] D. Ballou and G. K. Tayi, "Managerial Issues in Data Quality," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-95-03, February 1995.

[3] J. Bort, "Scrubbing dirty data," *InfoWorld*, vol. 17, no. 51, pp. 1-3, 1995.

[4] Cardinal Business Media, "Restoring Credibility to MIS," *MIDRANGE Systems*, vol. 8, no. 23, pp. 25, 1995.

[5] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377-87, 1970.

[6] N. Fenton, "Software Measurement: Why A Formal Approach," *Formal Aspects of Measurement*, Springer-Verlag, pp. 3-27, 1992.

[7] N. Fenton and S. L. Pfleeger, "Science and Substance: A Challenge to Software Engineers," *IEEE Software*, pp. 86-95, July 1994.

[8] C. Fox, A. Levitin, and T. Redman, "Data and Data Quality," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-95-05, February 1995.

[9] M. L. Fraim, "Decline Curve Analysis for Real Gas Wells with Non-Darcy Flow," in *Petroleum Engineering*: Texas A&M University, 1989, pp. 275.

[10] M. Hansen and B. Von Halle, "Visions of Power Rangers: Looking for a Successful Replication Strategy? Remember Halloween 1994," *Database Programming & Design*, vol. 8, no. 2, pp. 11-13, 1995.

[11] Y. U. Huh, F. R. Keller, T. C. Redman, and A. R. Watkins, "Data Quality," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-95-06, February 1995.

[12] A. Ishii, Y. Jang, and R. Wang, "A Qualitative Approach to Automatic Data Quality Judgment," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-93-02, May 1993.

[13] P. Kaomea, "Valuation of Data Quality: A Decision Analysis Approach," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-94-09, September 1994.

[14] H. Kon, J. Lee, and R. Wang, "A Process View of Data Quality," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-93-01, March 1993.

[15] H. Kon, S. Madnick, and R. Wang, "Data Quality Requirements Analysis and Modeling," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-92-03, December 1992.

[16] H. Kon, M. P. Reddy, and R. Wang, "Toward Quality Data: An Attribute-Based Approach," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-92-04, November 1992.

[17] T. A. Larsen, "A Study of Minor Groove Binding and Lexitropsin Target DNA Sequences by X-Ray Crystallography," in *Chemistry*. Los Angeles: University of California, 1990, pp. 161.

[18] K. C. Laudon, "Data Quality and Due Process in Large Interorganizational Record Systems," *Communications of the ACM*, vol. 29, no. 1, pp. 4-11, 1986.

[19] T. D. Leblanc, "Measuring the Quality of Educational Data in NWFP Pakistan," in *Education*: Harvard University, 1995, pp. 204.

[20] A. M. McGee, "Total Data Quality Management ... Zero Defects Data Capture," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-92-07, December 1992.

[21] R. C. Morey, "Estimating and Improving the Quality of Information in a MIS," *Communications of the ACM*, vol. 25, no. 5, pp. 337-42, 1982.

[22] V. Storey and R. Wang, "Modeling Quality Requirements in Conceptual Database Design," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-94-02, May 1994.

[23] D. Strong and R. Wang, "Beyond Accuracy: What Data Quality Means to Data Consumers," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-94-10, October 1994.

[24] B. Tandowski, "How clean is your data?," *Data Based Advisor*, vol. 10, no. 2, pp. 64-5, 1992.

[25] S. Y. Tu and R. Wang, "Modeling Data Quality and Context Through Extension of the ER Model," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-93-13, October 1993.

[26] USAF, *Air Force Manual 1-1, Basic Aerospace Doctrine of the United States Air Force*, vol. 1, 1992.

[27] Y. Wand and R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," Massachusetts Institute of Technology (MIT) Sloan School of Management, Cambridge, MA TDQM-94-03, June 1994.