A DATA QUALITY ENGINEERING FRAMEWORK

Dr. Donna M. Meyen
The MITRE Corporation
dmeyen@mitre.org

Dr. Mary Jane Willshire
Colorado Technical University
willshir@usa.net

Abstract

Databases that have multiple sources and contributors, multiple data managers, and which support a diverse set of users present significant challenges in sustaining the quality of the resident data. Research into data quality and data quality engineering arenas revealed that there was no strategic solution to the entire problem; a data quality engineering framework did not exist. This research draws upon the experience of one author working on a MITRE Corporation effort in support of the United States Air Force (USAF) to develop 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 were viewed with concern as to their quality. For example, the accuracy and timeliness of the data were suspect because the data had not been updated since their initial collection.

In [19] we presented in general terms a Data Quality Engineering Framework (DQEF), where various single solutions can be integrated in order to effect data quality process and product improvement. This paper presents the results of applying the DQEF to the USAF environment. A detailed presentation of this effort is given as an example of how one could use the framework and customize it for an organization/enterprise. Conclusions and recommendations are offered, along with areas where further investigation is indicated.

1. Introduction

Databases that have multiple sources and contributors, multiple data managers, and which support a diverse set of users present significant challenges in sustaining the quality of the resident data. Research into the identification of candidate data quality engineering processes revealed that various facets of data quality had been investigated and reported upon in numerous fields [4, 9, 16, 17, 18]. In addition, mathematicians, information management experts, and computer scientists have offered numerous ways in which to define [8, 13], analyze [14, 15], and improve the quality of the data [1, 2], regardless of the format, media, or content. Both automated tools [3] and formal methods for data comparison [11] have been made available. However, none of the mentioned approaches organized these partial solutions in a coherent fashion to provide an overarching framework in attempting to permanently solve a data quality problem. In short, no one had developed an entire data quality engineering process or program from beginning to end.

Research was initiated that drew upon experience of one author working on a MITRE Corporation effort in support of the United States Air Force (USAF) to develop a top-level process and method for understanding, managing, and directing the architectures, interoperability, and evolution of complex systems of systems. When this effort began, data regarding these systems were collected. Various users of the database raised concerns regarding the quality of the collected data. For example, the accuracy and timeliness of the data were suspect because the data had not been updated since their initial collection.

This paper is a report on research that developed a general Data Quality Engineering Framework (DQEF) that can be used to define, analyze, and provide guidance to improve data quality. Section 2 discusses the general method for developing and applying the DQEF. Section 3 provides the results of applying the DQEF to the USAF environment. Section 4 summarizes the results of the research.

2. The General Method

Our method to develop and prove the viability of the Data Quality Engineering Framework (DQEF) consisted of six steps, each of which is described in more detail in this section. The first step developed the general framework (DQEF) that would be used to define, analyze, and provide guidance to improve data quality. The second step defined measures of success in developing and applying the DQEF. The third step tailored and applied the DQEF to a specific data environment. Describing and analyzing the results of applying the tailored DQEF comprised the fourth step, while drawing conclusions from the application of the DQEF was the fifth step; these latter two steps determined the viability of the DQEF. Recommendations were formulated and presented as step six.

In the first step, we used a wide range of source experts in the area of data quality and best engineering judgment to develop the DQEF 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. A major feature of the DQEF is that it specifically addresses the temporal aspect of data quality environments. There are four major steps in the DQEF: define a Data Quality Engineering Model (DQEM); define the relevant data quality attributes; analyze the data quality attributes; and provide data quality improvement guidance. The DQEF is described in detail in [19].

The second step in developing the DQEF was to define measures of success. These were the proof criteria used to judge whether or not the DQEF was successful. These were also described in [19] and included:

1. The DQEF was successfully used as a basis to **define** data quality attributes specific to a particular data environment and users' needs. This was a subjective decision, and binary in nature; either the relevant data quality attributes were identified and defined or they were not.
2. The DQEF was successfully tailored to facilitate the measure and **analysis** of the quality of a specific set of data. This was a qualitative decision, and also binary. Again, either the framework supported measurement and analysis activities or it did not.
3. The DQEF provided guidance in identifying methods to **improve** the quality of data in a selected case. A comparison was made between an initial measurement and a second measurement, for all types of measurements recorded. Qualitative measures were required to show a one-level improvement in at least 50% of qualitative attributes after the implementation of remedial actions. Quantitative measures were required to show at least a 5% improvement in at least 50% of quantitative attributes after the implementation of remedial actions.

The third step was to execute the DQEF using selected portions of an existing database as a test case in order to verify the DQEF. The functions requiring the services of the associated database were described, and the services provided by the database to these functional areas were also documented. The logical structure of the database and the appropriateness of applying the DQEF to this database were identified. This provided the context for tailoring the

DQEF based on descriptions in the DQEM. Available alternatives for each DQEF task were evaluated in light of this specific environment.

In step four, the results of applying each of the DQEF's activities were analyzed in detail in order to validate the framework, demonstrating that the DQEF is a viable framework. Comparison of data quality measurements was performed, along with an analysis of the root causes of any discovered problems. Remedies were incorporated, and the results of their implementation recorded and analyzed.

Conclusions were drawn in step five of the method. A determination was made as to whether the success criteria for developing the DQEF were met. The analysis of the quantitative and qualitative measurements provided the basis for all conclusions.

The sixth and final step of the method was to offer recommendations indicated as a result of executing the previous five steps. Recommendations were made concerning the general usage of the DQEF, as well as extensions that could be made to the framework. Recommendations regarding specific areas within each phase of the DQEF, such as requirements elicitation, assessment methods, and automated tools, were also suggested.

The next section provides a brief case study where this method was applied in the context of the Air Force (AF) Command, Control, Communications, Computers and Intelligence (C^4I) environment.

3. A Specific Application of the DQEF

One of the authors has been involved in an effort to assist the USAF in understanding, managing, and directing the evolution, interoperability and architectures of its complex systems of systems. In the course of this work, the AF C^4I Database was constructed containing the data describing these systems. A detailed overview of the AF C^4I data environment (the AF C^4I DQEM) spanning the time period between 1995 and 1998 is presented in [19]. Section 3.1 describes the remainder of the tailored DQEF, that is, the data quality requirements, the methods chosen for measurement and analysis, and the remedies selected for incorporation into the subject data environment. Section 3.2 provides an analysis of tailoring the DQEF to the AF C^4I Database environment. Sections 3.3 and 3.4 present conclusions and recommendations respectively.

3.1 The AF C^4I Database Data Quality Engineering Model (DQEM)

3.1.1 Definition of Data Quality Parameters for the AF C^4I Database

Once we established our baseline DQEM [19], we defined the data quality parameters peculiar to this environment. To do so, we selected data quality parameters based on the users' requirements such as data criticality, functional context, and expert opinions.

In a recent survey, a list of 179 data quality attributes was compiled and categorized [23]. This comprehensive list, presented in Figure 1, provided the basis for selection of data quality attributes which were pertinent to the AF C^4I Database environment. As the selections were made, definitions for each data quality attribute were recorded and a rationale for each choice documented. Both qualitative and quantitative attributes were selected.

The attributes selected by the primary user of the data were heavily biased towards data integration, data analysis, and data presentation because the primary user must have the ability to summarize the data. The secondary

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 1. Generic Data Quality Attributes [23]

user selected different attributes since the data is used to determine whether logical and/or physical connections exist between nodes. The secondary user was also concerned with the expense and cost of collecting these data. A third user chose yet other attributes, reflecting the level of detail necessary to determine system compatibilities. A fourth user, in its role as Central Maintenance Organization (CMO), was concerned with the maintainability of the data, and selected attributes reflecting this aspect.

The attributes chosen by each group of users were translated into user requirements, expressed as questions, which assisted in identifying the way the attributes were perceived by each user. Each requirement was then mapped to

one of four Major Categories using Strong's guidelines [23]: accuracy, accessibility, relevance, and representation. Figure 2 shows a sample of the mapping from attributes to Major Categories.

Attribute	User Requirement	Sub-Category	Major Category
Ability to Download	Can the data be downloaded to a local host?	accessibility	Accessibility
Accuracy	Are the data accurate?	accuracy	Accuracy
Adequate Volume	Are there enough/too much data?	aaod[1]	Relevance
Age	How old are the data?	timeliness	Relevance
Aggregatability	Can the data be summarized easily?	ease of ops	Accessibility
Amount of Data	Are there enough/too much data?	aaod	Relevance
Authority	What is the source of the data?[2,3]	reputation	Accuracy
Believability	How believable are the data?	believability	Accuracy
Breadth of Data	Do the data cover enough areas? How complete are the data?	completeness	Accessibility
Clarity	Is the meaning of the data clear?	ease of understanding	Representation
Clarity of Origin	What is the source of the data?	reputation	Accuracy
Completeness	Do the data cover enough areas? How complete are the data?	completeness	Accessibility
Comprehensiveness	Do the data cover enough areas? How complete are the data?	completeness	Accessibility
Compressibility	Can the data be summarized easily?	ease of ops	Accessibility
Consistency	Are the data consistent?	rep consis	Representation
Context	Are the data relevant?	relevance	Relevance
Correctness	Are the data accurate?	accuracy	Accuracy
Corruption	Are the data accurate?	accuracy	Accuracy
Current	How old are the data?	timeliness	Relevance
Dependability	What is the source of the data?	reputation	Accuracy
Detailed Source	What is the source of the data?	reputation	Accuracy
Ease of Comparison	Can the data be summarized easily?	ease of ops	Accessibility
Ease of Correlation	Can the data be summarized easily?	ease of ops	Accessibility
Ease of Retrieval	Can the data be downloaded to a local host?	accessibility	Accessibility
Ease of Understanding	Are the data consistent?	interpretability	Representation
Ease of Use	Can the data be summarized easily?	ease of ops	Accessibility
Easy to Question	Can the data be summarized easily?	ease of ops	Accessibility
Enlightening	Do the data add value to the operation?	value added	Relevance
Error-Free	Are the data accurate?	accuracy	Accuracy

Figure 2. Sample Mapping from Attribute to User Requirement to Categories

After the data quality attributes were assigned to Major Categories, an analysis was performed to determine which data quality attributes appeared most frequently in the user selections. For each user, the data (see Figure 2) were sorted according to Major Category. The number of items contained in each User Requirement Category were then counted. Figure 3 summarizes the findings for all users. The results indicated where the focus of further research should occur. The Major Categories of Accuracy and Accessibility accounted for over 62% of the users' requirements. Within these Major Categories, the Sub-Categories of "ease of operations" and "completeness" accounted for over 70%

[1] appropriate amount of data

[2] For the purposes of this exercise, the source of the data is assumed to be the Subject Matter Expert (SME) of the data

[3] For the purposes of this exercise, the origin of the data is assumed to be the source of the data

of Accessibility attributes, while "accuracy" and "reputation" accounted for over 80% of Accuracy attributes in the current AF C^4I Database environment. Figures 4 and 5 depict these results.

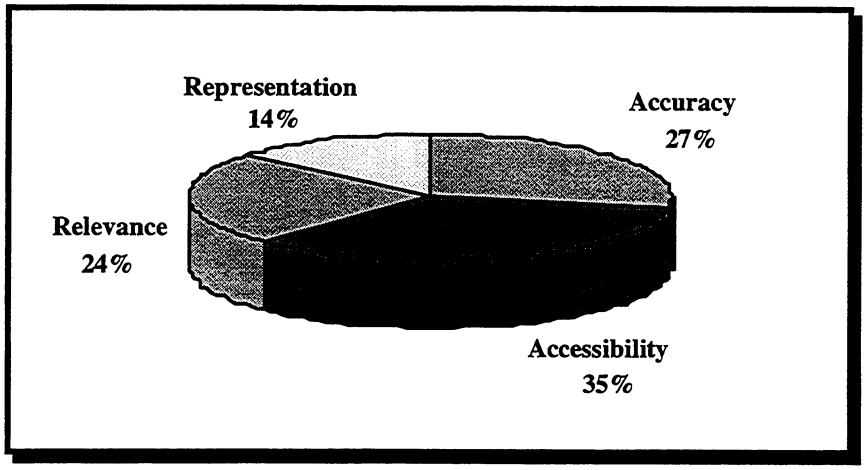

Figure 3. Summary of Data Quality Major Categories - All Users

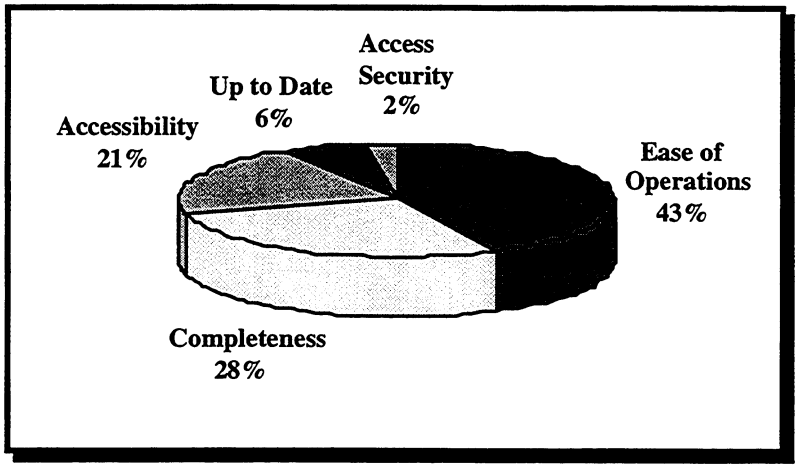

Figure 4. Accessibility Sub-Categories -- All Users

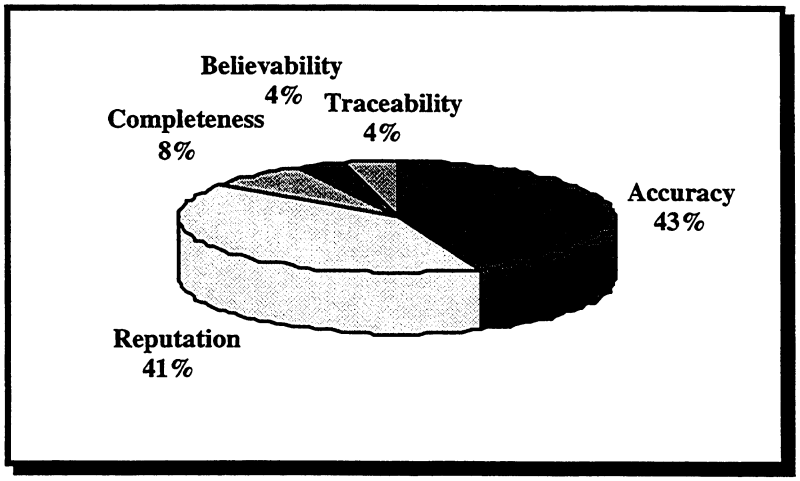

Figure 5. Accuracy Sub-Categories -- All Users

3.1.2 Collection, Measurement, and Analysis of Data Quality Attributes

The next effort was to select the methods used to collect and measure data. The attributes selected by the process described above were categorized as either external or internal as defined by Fenton in [6]. In addition, a measurement scale as described by Fenton in [7] was used as a framework within which attributes were measured and compared. If the attribute was considered external, then internal attributes that could be measured to indicate the extent of this specific external attribute were identified. All attributes were then assigned the type of scale to be used in their measurement. All data collection activities were accomplished manually.

The relevant objective and subjective quality indicators that have been suggested in the literature [21, 22, 25] as indicative of the external attributes upon which this research focused are listed in Figure 6. The first ten objective indicators (internal attributes) listed in Figure 6 are currently used in several automated tools to assist in measuring the structural and representational quality of a database. The remaining two indicators are subjective internal attributes used either manually or in concert with an automated tool to judge the actual content and fitness of use of a database. Subject Matter Experts (SMEs) and the database users themselves evaluated the quality of the data subjectively and provided a High, Medium, or Low rating.

Objective Attribute	**AF C^{4}I Database Measurement**
1. Range of Values	0.98 percent out of tolerance
2. Domain Values	0.98 percent out of tolerance
3. Cyclic Redundancy Check	Not necessary in current configuration; planned for future
4. Units	.06 percent out of tolerance
5. Business Rules	Do not exist
6. Consistency Checks	285 diagram/data disconnects; 229/277 inter-mission disconnects
7. Standard Definitions	Do not exist
8. Metadata	Exists for all fields in database
9. Presence of Value	28% empty cells
10. Linking	Does not exist
Subjective Attribute	**AF C^{4}I Database Measurement**
20. SME Evaluation	Accuracy = L; Reputation = M^+; Completeness = M^-
21. User Evaluation	Completeness = M^-; Ease of Ops = M

Figure 6. Summary of Initial Measurements of AF C^{4}I Database

The result of measuring the AF C^{4}I Database using these methods is also presented in Figure 6. In the Completeness category, approximately 28% of the AF C^{4}I Database cells contained no data. Metadata exists and is documented for all AF C^{4}I Database fields/cells. Documentation describing the Range of Values (ROV) and Domain Values for all fields also exists [10]. To measure ROV and Domain Values, it was necessary that a value be present in the cell. The value was then tested for being within the ROV and the Domain tolerances defined by the metadata. Of those present, almost 1% of the data values violated the tolerance levels defined for the AF C^{4}I Database. Appropriate units for expressing the data have not been defined nor are present in the AF C^{4}I Database for approximately .06% of the

database. This is in addition to those values that violate the standards for ROV and Domain values. Standard Data Element Definitions do not exist for the AF C^4I Database.

A consistency check was performed between inter-mission source and destination pairs, resulting in the discovery that 83% of the source/destination pairs were not consistently represented between missions. Cyclic Redundancy Checks (CRCs) are not employed in the data life cycle because all data transfer takes place via floppy disks and the operating system software employed to copy the data to the disks uses a standard integrity check. In addition, business rules have not been documented stating relationships between the data elements. No linking exists between tables for correlation between data sources and data origins.

Once the objective and subjective measurements were taken, various methods of manipulating the measurements to assess the quality of the AF C^4I Database were chosen. The selection criteria included suitability to the application, suitability to the data, and the ability of the method to yield a meaningful result. A major selection criterion was that the chosen methods approached the data environment from orthogonal aspects, so that examples of assessing a database based on differing viewpoints could be demonstrated. Four methods were chosen and are described in Sections 3.1.2.1 through 3.1.2.4.

3.1.2.1 Decision Analysis

One assessment method chosen for this research was the decision analysis approach described by Kaomea in [12]. This method computes the value of data qualities in a given decision scenario. It is envisioned that the data contained in the AF C^4I Database will be considered when making decisions about new system acquisitions and capabilities, so this method was deemed relevant.

The decision analysis method first computes the value of a data quality attribute as the product of the values of related sub-attributes. Therefore, for example, the value of the attribute "accuracy" can be computed as the product of the values of "source accuracy", "source credibility", and "data clarity." A decision tree which models the situation is then built, and the computed data quality values are inserted into the decision tree. A sensitivity analysis is then performed by varying each sub-attribute value, one at a time, and comparing the results of the outcomes.

This method was applied to the AF C^4I Database in terms of the primary user requirements. The example situation is where the primary user must decide whether a new system acquisition should be allowed to continue, based on whether the system will perform at acceptable levels of interoperability with existing systems. Without any data to impact the decision, the decision tree formulated is shown in Figure 7. Equally likely events are acceptable and non-acceptable levels of interoperability. If the system possesses an unacceptable level of interoperability and is terminated, nothing is lost. If the system possesses an acceptable level of interoperability and is terminated, then 20 million dollars (M) are at risk. If the system possesses an unacceptable level of interoperability and is allowed to continue, then those same 20M are at risk, as well as sunk costs of the program -- an additional 10M -- for a total risk of 30M. If the system possesses acceptable levels of interoperability and is allowed to continue, then nothing is lost. These are represented as outcomes o1 through o4 respectively on Figure 7.

Inserting the quality attributes of data availability and data accuracy into the primary user decision scenario yielded the decision tree in Figure 8. The AF C^4I Database data availability was measured and is reported in Figure 6

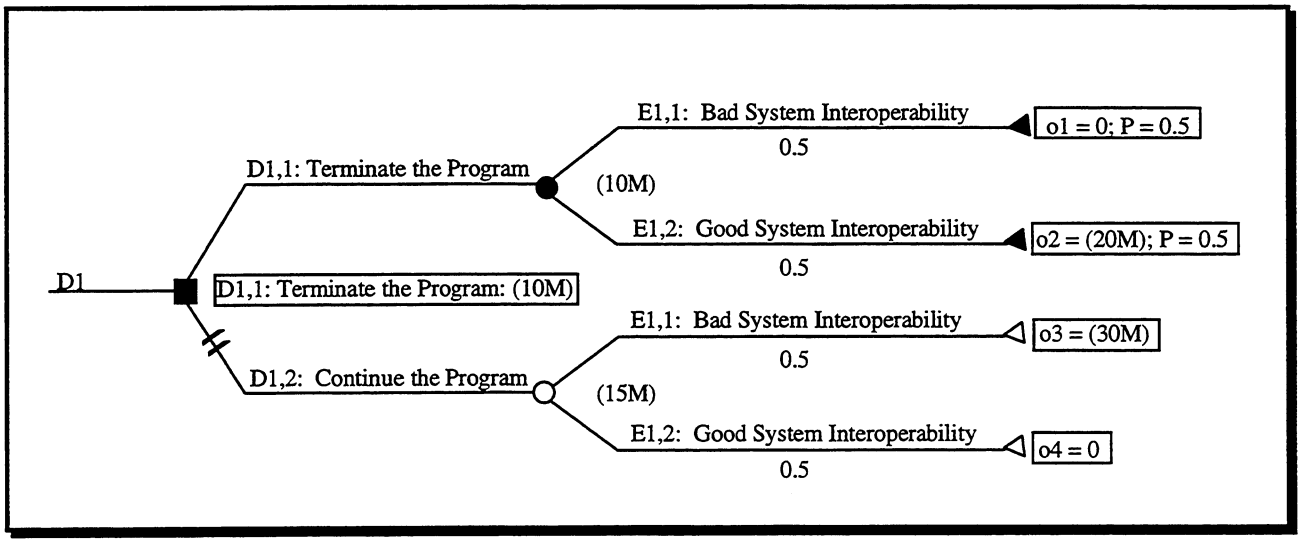

Figure 7. Initial Decision Tree for the Primary User's Mission

as 72%. Data accuracy is a product of the measures 1 through 8 and 20 on Figure 6. Respectively, the measures for 1 through 8 are:

1. 0.990 (Range of Values)
2. 0.990 (Domain Values)
3. 1.000 (CRC)
4. 0.995 (Units)
5. 1.000 (Business Rules)
6. 0.170 (Consistency)
7. 1.000 (Standard Definition)
8. 1.000 (Metadata)

Item 20 was computed as follows: a rating of Low received a value of 1, a rating of Medium received a value of 2, and a rating of High received a value of 3, so that in the subjective accuracy Sub-Categories, a possible total of 15 was attainable. Each subjective accuracy Sub-Category in Figure 6 received a rating of 1 for a total of 5, yielding a measurement of 33%. Sub-Categories 3, 5, 7, and 8 hold values of 100%; they do not affect the product and were omitted from further calculations. The values used to perform the AF C^4I Database primary user decision analysis are reflected in the first row of Figure 9.

The product of these measurements was taken to arrive at the computed value of the AF C^4I Database data accuracy value of 5%. This value, as well as the value for availability (72%), was inserted into the decision tree and resulted in the decision to terminate the program, thereby minimizing losses at approximately 3M. Figure 10 reflects this decision tree. One at a time, the values of each of the accuracy Sub-Categories were brought to perfection, and the results were re-computed. The second through sixth rows of Figure 9 reflect these results as well. Note that as data accuracy increased, so did the amount of loss. However, losses remained within acceptable limits; it appeared that none of the accuracy Sub-Categories had a great effect on the outcome.

Kaomea's method was then used to conduct a sensitivity analysis on the availability attribute. The availability measure was given values of 90% and 10% respectively, and the outcomes were re-computed at each of the accuracy

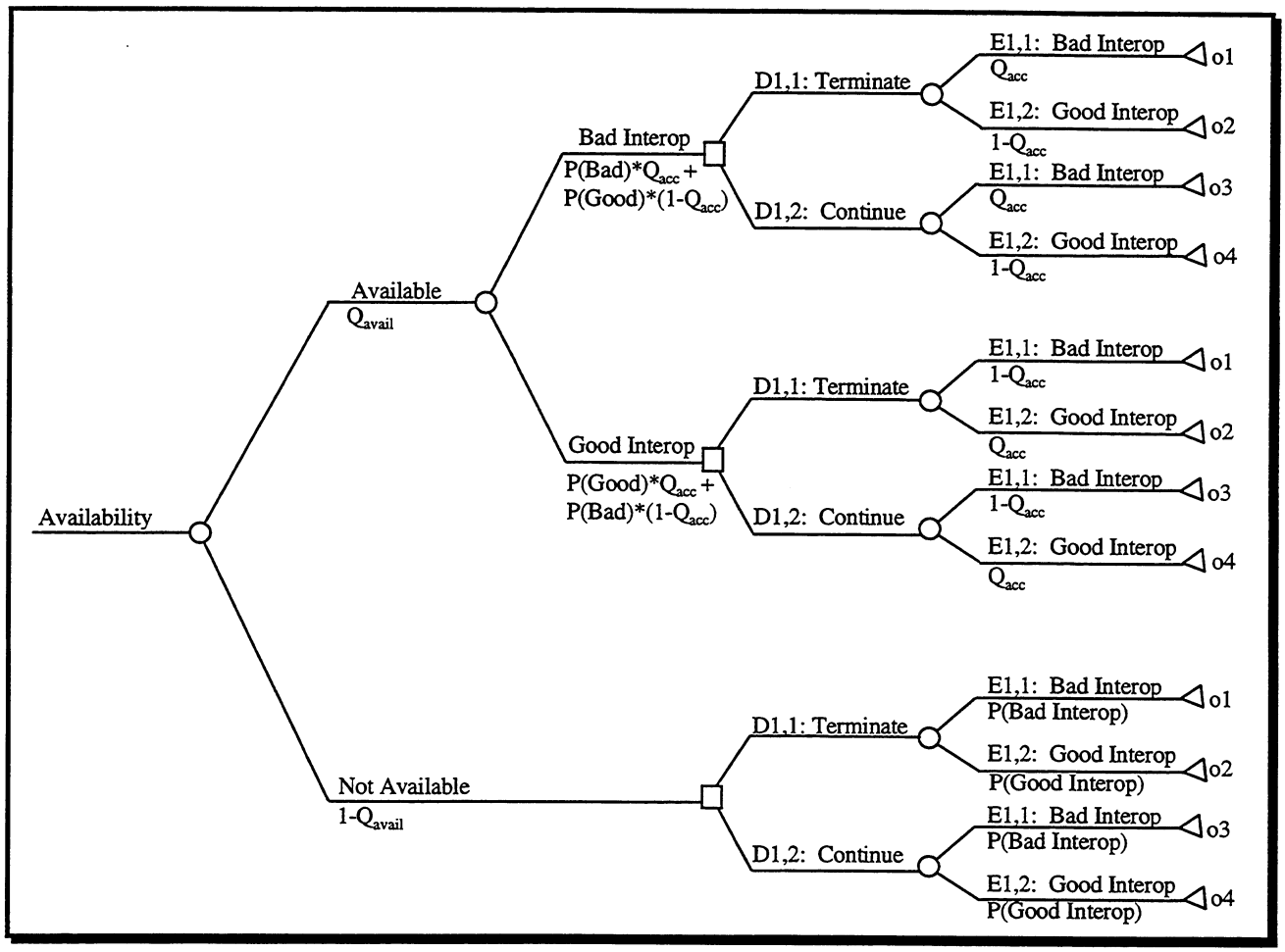

Figure 8. Primary User Decision Tree with Quality Attributes Inserted

Case	Range of Values	Domain Values	Units	Consistency	SME Evaluation	Data Accuracy	Outcome
Original	0.99	0.99	0.995	0.17	0.33	0.05	-3.16M
Range Perfect	1	0.99	0.995	0.17	0.33	0.05	-3.16M
Domain Perfect	0.99	1	0.995	0.17	0.33	0.05	-3.16M
Units Perfect	0.99	0.99	1	0.17	0.33	0.05	-3.16M
Consistency Perfect	0.99	0.99	0.995	1	0.33	0.32	-5.10M
SME Evaluation Perfect	0.99	0.99	0.995	0.17	1	0.17	-4.02M

Figure 9. Accuracy Sub-Categories Impact on Primary User's Decision Analysis Outcome

levels. The results are reflected in Figure 11. Note that in this case, availability of the data appears to have a significant impact on the outcome, indicating that emphasis should be placed on improving the availability of the data, at least for the primary user's requirements.

3.1.2.2 Entity-Relationship Extension

The second method chosen to measure the quality of data was to extend the Entity-Relationship (ER) model to include quality attributes [14, 24]. The original theory stated by Kon [14] defined a data quality parameter as a

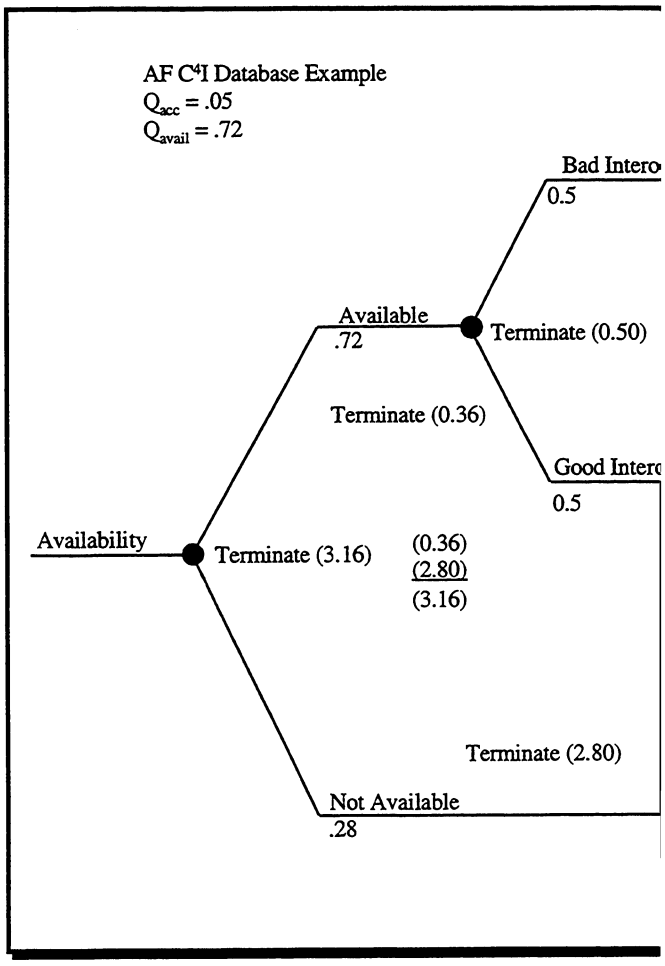

Figure 10. Primary User Decision Tree

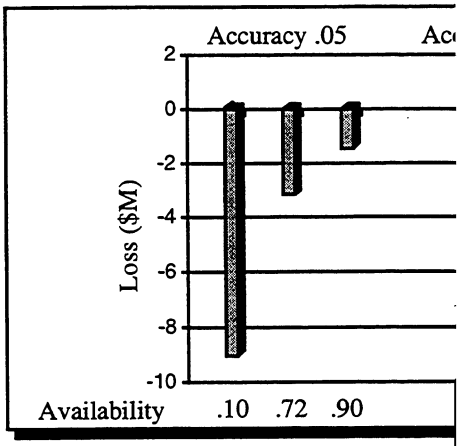

Figure 11. Availability Impact on Prim

"qualitative or subjective dimension by which a user evalu attributes defined by Fenton in [6]. Kon extended the ER m to the attributes of the entities defined in the relations. D

parameters, and would be assigned values. The user would then be able to judge the quality of the data concerning the entity based upon the value of these data quality parameters. Since the AF C^{4}I Database was being managed by a relational database management system, and already included some data quality attributes, this method was an appropriate choice for our use. In addition, this method allows users to assess the quality of the database, something which the Decision Analysis method does not address.

The AF C^{4}I Database has no cell-level tags denoting data quality attributes. There are, however, row-level attributes which assist the user in determining the quality of the data. For each table in the AF C^{4}I Database, the presence of row-level tags was determined, as well as any values within the respective tags. Based on the presence of tags, and the value contained within those tags, it was possible to assign a level of quality to the AF C^{4}I Database. These results are shown in Figure 12. Less than 15% of the total number of data quality attribute cells in the AF C^{4}I Database were meeting an acceptable value threshold. At best, only 40% of non-empty data quality attribute cells were meeting an acceptable value threshold.

The research also experimented with a slightly different modification to Kon's method. It was determined in Section 3.1.1 that the two Major Categories of Accuracy and Accessibility would be the focus of this research. The respective Sub-Categories being emphasized within these Major Categories were accuracy and reputation (for Accuracy), and completeness and ease of operations (for Accessibility). For each Sub-Category, cell-level tags were developed and defined which would assist the user in evaluating the quality of the data within the AF C^{4}I Database. These tags are listed in Figure 13.

Table	No. of Cells	No. of Non-Empty Cells	No. Meeting Threshold	Pct Overall Meeting Threshold	Pct Non-Empty Meeting Threshold
inter95	13170	4774	2835	21.53	59.38
inter05	6030	2613	84	1.39	3.21
act_item	618	30	0	0.00	0.00
comm_sys	390	200	101	25.90	50.50
issue	3246	106	96	2.96	90.57
Grand Totals	23454	7723	3116	13.29	40.35

Figure 12. AF C^{4}I Database Cells Meeting Data Quality Attribute Thresholds

Accuracy		Accessibility	
Accuracy	**Reputation**	**Completeness**	**Ease of Operations**
SME Name/ Org	Traceability to Origin of Data	Level of Detail Indicator	Date of Last Update
Validated by SME Date	Name of Source/ Org and Contact Information	Number of Empty Cells	Date of Next Update
SME Validation Exp Date			Data Formats/Standards

Figure 13. AF C^{4}I Cell-Level Data Quality Attribute Tags

Under the Accuracy Sub-Category, the SME name and/or organization identifies the validator of the data. The most recent validation date gives the user an idea of the age of the data, while the expiration date indicates to the user that the data are not to be trusted after that date. In the Reputation Sub-Category, the identity of the origin of the data, or a way to track the origin of the data, would be provided by the traceability to the origin of the data. If the origin is different from the source, then the name and/or organization of the source would also be indicated. Under Completeness, the level of detail indicator could be chosen from a pre-defined, standard list. Each level would be thoroughly described and standardized. The number of empty cells gives an indication of the completeness of the row of data. Under Ease of Operations, the date fields (last update and next expected update) give an indication to the user and/or SME of the necessity to update. The data formats and standards adhered to by the data would give an indication of the ability to transport data across different platforms. Other indicators could be added as necessary.

The list presented in Figure 13 was then compared to the actual table structures and contents in the AF C^4I Database. The results are presented in Figure 14. A check mark denotes that the structure was already present in the AF C^4I Database. A second check mark indicates that the structure was present, with data values present. An empty entry indicates that the structure was not present.

Table	SME Name /Org	SME Validation Date	SME Validation Expiration Date	Trace-ability to Origin	Name of Source	Level of Detail Indicator	No. of Empty Cells	Date of Last Update	Date of Next Update	Data Formats/ Standards
inter95		√			√√			√√		
inter05		√			√√			√√		
ac_radio										
acronyms										
act_item								√		
ais										
comm_sys					√√			√√		
comm_ntwk										
elements										
glossary										
issue				√√	√√			√√		
person										

Figure 14. Data Quality Attribute Row-Level Structures

3.1.2.3 Mathematical Estimation

A third method is based on mathematical foundations, as in Morey [20]. This work was based on the data life cycle in a transaction-based system. Morey defined three key measures: the transaction reject rate (r), the intrinsic transaction error rate (e_T), and the stored Management Information System (MIS) record error rate (e_M). Morey went on to develop estimation formulas for error rates (r, e_T, e_M) and how to apply them to an analysis of the effectiveness of error reduction mechanisms. Since the data life cycle discussed in Morey's work closely paralleled that of the DQEM previously developed, his work was integrated into this research. This assessment method concentrates on predicting

the accuracy of a database, which is different from assessing the sensitivity to a particular attribute or assessing the overall quality of a database, giving yet a third orthogonal view of the quality of the database.

The data life cycle discussed by Morey in [20] is depicted in Figure 15. Morey defined several parameters that are factors in the mathematical model:

1. P = the probability that an erroneous transaction is properly rejected by an edit check.
2. $1 - P$ = the probability that an erroneous transaction is not properly rejected by an edit check (Type I error).
3. P' = the probability that a correct transaction is improperly rejected an edit check, thereby delaying proper updating of the record (Type II error).
4. T = a non-negative random variable representing the time interval or spacing between transactions for a given record.
5. μ_T = mean of the intertransaction times

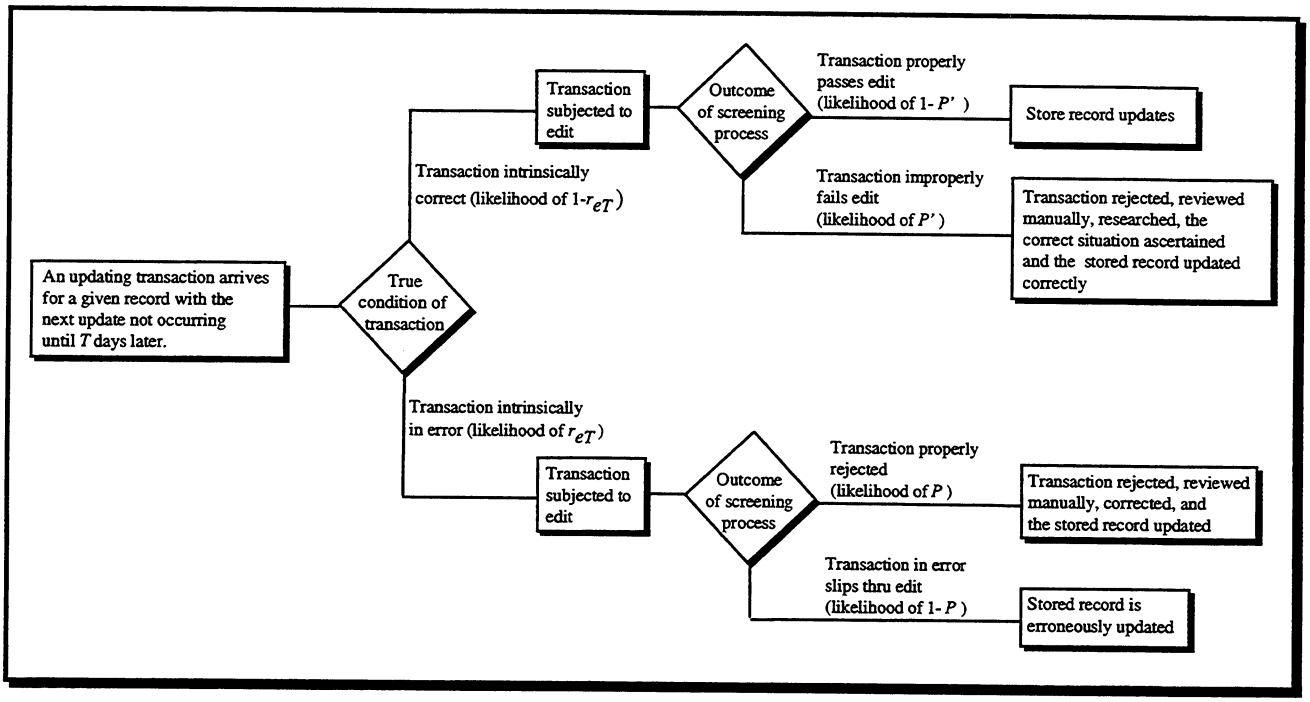

Figure 15. Data Life Cycle Described by Morey [20]

6. C_1 = minimum processing time; the elapsed time from when a transaction is submitted to the system until it updates the record.
7. C_2 = additional processing time delay, over and above C_1, to manually review and correct transactions which (i) were in error, and (ii) were properly rejected by the edit checks. Morey assumed that the manual review perfectly resolved all discrepancies and correctly updated the stored record.
8. C_3 = additional processing time over and above C_1 to manually review and allow to enter into the system any intrinsically correct transactions which were rejected by the edit checks. Morey assumed

the reviewer was able to ascertain the correct situation so that the stored record was updated accurately.

9. r_e = empirical estimate of the transaction reject rate (simply the fraction of transactions in any sample that were rejected by the edit checks).

Morey defined the intrinsic transaction error rate (when $P > P'$) as:

10a. $e_{rT} = 0$ if $r_e > P'$

10b. $e_{rT} = (r_e - P') / (P - P')$ if $P' \leq r_e \leq P$

10c. $e_{rT} = 1$ if $r_e > P$

And finally, Morey defined the stored MIS record error rate as:

11. $e_{rM} \geq e_{rT}(1 - P) + [C_1(1 - e_{rT})(1 - P') + (C_1 + C_2)e_{rT}P + (C_1 + C_3)(1 - e_{rT})P']/\mu_T$

Based on the measurements listed in Figure 6, the following values were substituted into Morey's equations. P was set to 99% because the AF C^{4}I Database contained a little less than 1% out-of-tolerance values for ROV and Domain values. P' was set to .5% because there are very few edit checks performed by the data management software; most of the values input into the AF C^{4}I Database pass all edit checks. Based on P of 99%, 1 - P is 1%. T was set to 180 (days) because two releases per year (therefore two updates per year) are scheduled. μ_T was set to 180 because that is the mean of intertransaction times (≈365/2). C_1 was set to 7 days, while C_2 was set to 3 days and C_3 to 5 days, based on the empirical data for the last two releases of the database. r_e was set to 2%, because only 2% of the limited edit checks are rejected. P was greater than P', and $P' < r_e < P$. Therefore, equation 10b was used to compute the intrinsic transaction error rate, e_{rT}:

$e_{rT} = (0.02 - 0.005) / (0.99 - 0.005) = 0.015$

Once the intrinsic transaction error rate was computed, the stored MIS record error rate could be computed using equation 11:

$e_{rM} \geq 0.015(1 - 0.99) + [7(1 - 0.015)(1 - 0.005) + (7 + 3)(0.015)(0.99) + (7 + 5)(1 - 0.015)(0.005)] / 180$

≥ 0.03942

Using Morey's method, then, it was estimated that at least 4% of the stored data values in the AF C^{4}I Database were in error.

3.1.2.4 Reduction-Based Data Quality Calculus

The final method was based on Ishii in [11]. This method is qualitative and is referred to as a "reduction-based data quality calculus." Since this method focused on qualitative attributes of data obtained from several different data sources, and the AF C^{4}I Database contains data from several sources, this was judged to be an appropriate method. This method dealt with subjective measurements, while the other three methods dealt with objective measurements. Ishii's method derives an overall data quality value based on relationships among the data quality attributes as dictated by the users' context. The following paragraphs present the scenario and basic measurements adopted for the computation of the AF C^{4}I Database quality using this method.

As part of their mission, the primary user may attempt to excise non-interoperable system acquisitions from the fiscal budget. They may want to determine if the data in the AF C^{4}I Database are believable enough to contribute to

these decisions. Factors affecting the believability of data are temporal effect, reputation of the data source, and accuracy of the data present in the database (for example). Therefore, the quality parameters (QP) that the primary user may employ as a gauge are:

QP = {Temporal-Effect, Accuracy, Reputation}.

For the purposes of this example, it was assumed that Temporal-Effect dominated Reputation and Accuracy, and that Reputation dominated Accuracy. In other words, the primary user would rather have current data over and above a reliable source or accurate data. Therefore, the set of Dominance Relations (DR) consisted of the following dominance relationships between Temporal-Effect, Reputation, and Accuracy:

Temporal-Effect dominates Accuracy for all {Temporal-Effect := Tolerable |Moderate | Intolerable} and all {Accuracy := High | Medium | Low}

Temporal-Effect dominates Reputation for all {Temporal-Effect := Tolerable | Moderate | Intolerable} and all {Reputation := High | Medium | Low}

Accuracy dominates Reputation for all {Accuracy := High | Medium | Low} and all {Reputation := High | Medium | Low}.

Based on the SME estimates provided in Figure 6, Temporal-Effect was assigned a value of Intolerable, Reputation was assigned a value of High, and Accuracy was assigned a value of Medium.

In computing the believability of the data in the AF C^{4}I Database, the instantiated quality-merge statement was reduced by algorithm Q-Reduction:

1. $\Omega \leftarrow$ {Temporal-Effect := Intolerable, Accuracy := Medium, Reputation := High}
2. $\Omega \leftarrow \Omega$ — {Reputation := High} since the dominance of Temporal-Effect and Accuracy over Reputation is asserted in DR. This leaves Ω = {Temporal-Effect := Intolerable, Accuracy := Medium}
3. $\Omega \leftarrow \Omega$ — {Accuracy := Medium} since the dominance of Temporal-Effect over Accuracy is asserted in DR. This leaves Ω = {Temporal-Effect := Intolerable}.

This reduction implies that the primary user is only concerned about the temporal degradation of the AF C^{4}I Database. As a result, the Q-Reduction algorithm returned a Believability value of Intolerable.

3.1.3 Analysis and Interpretation of Results

This section reports on the remedies selected for incorporation into the AF C^{4}I Database DQEM.

Based on the initial data quality measurements, an analysis and interpretation task was initiated to identify data quality problem areas and their probable/possible/definite cause(s). The objective was to recommend modifications to the DQEM which would

1. prevent further pollution of the AF C^{4}I Database by eliminating/reducing causes of low-quality data
2. provide each class of user with an indication of the quality of the data/database
3. provide a means to identify and repair unacceptable values in the database, and to disseminate these values

4. possess an optimal cost/benefit ratio.

A combination of remedies was selected for incorporation into the data life cycle. These remedies are described in the following paragraphs.

The implementation of the Decision Analysis approach resulted in a clear indication that availability of data is where remedial action should be applied. Because the data life cycle appeared to be extremely sensitive to the availability of data, the remedial action to be taken, then, would be to complete the population of the AF C^4I Database.

The extension of the AF C^4I Database structure to include quality attributes resulted in two clear observations. The first, indicated by Figure 12 showed that there were quality attribute indicator structures already present in the AF C^4I Database. However, 67% of these cells were empty. Of the cells containing values, only 40% met acceptable thresholds for consumer use. The second observation was the AF C^4I Database structure contained very few data quality indicators, as shown in Figure 14. It was concluded from these two observations that quality indicators should be added to the AF C^4I Database structure and that these indicators should be completely populated. An update of the data was also indicated by these observations.

Using the mathematical estimation technique resulted in an indication that at least 4% of the data in the AF C^4I Database were incorrect. A sensitivity analysis of the various factors affecting the result was conducted, similar to that performed for the implementation of Kaomea's decision analysis approach. Each of the parameter values was varied and the results were recorded. The results of this analysis are presented in Figure 16. These results indicated that if the update rate was increased to four times a year, while the submittal rate was decreased to one day, a 1% rate in incorrect data would result.

The data quality calculus technique resulted in an indication that the users strongly prefer current data over and above accurate or reliable data. This would reinforce the suggestion to update the AF C^4I Database more frequently.

The next section presents the last three steps of the general method of proving the hypothesis: analyze results, draw conclusions, and make recommendations.

3.2 Analyze Results

The AF C^4I Database data life cycle model was revised to incorporate the remedies discussed in Section 3.1.3. A second assessment, using the same four methods selected for making the initial assessments, was then performed. An analysis of the impact of the adjusted data life cycle on the quality of the data was undertaken to reveal the improvement/decline in the quality of the database. In other words, the verification and validation of the DQEF took place, where it was demonstrated that DQEF is a viable tool to define, analyze, and improve data quality for this environment. Comparison of data quality measurements was made; the results of the analysis are presented in Figure 17; they were compared to the Measures of Success when determining the viability of the DQEF.

Figure 17 indicates that all four assessments showed significant improvement after the data life cycle model was adjusted to incorporate the remedies discussed in Section 3.1.3. All objective measurements improved by at least 25%, with the largest gain at almost 1900%. All subjective measurements improved by at least one level, except Reputation, which remained at the highest level. This served to verify that the DQEF produces accurate results and

Case	P	1-P	P'	μ_T	C_1	C_2	C_3	r_e	e_{rT}	e_{rM}
Original	0.99	0.01	0.005	180	7	3	5	0.02	0.015228426	0.039423294
Increase P to .994	0.994	0.006	0.005	180	7	3	5	0.02	0.015166835	0.039364397
Increase P to 1	1	0	0.005	180	7	3	5	0.02	0.015075377	0.039276941
Decrease P' to .004	0.99	0.01	0.004	180	7	3	5	0.02	0.016227181	0.039421907
Decrease P' to 0	0.99	0.01	0	180	7	3	5	0.02	0.02020202	0.039416386
Decrease T to 90	0.99	0.01	0.005	90	7	3	5	0.02	0.015228426	0.078694303
Decrease T to 30	0.99	0.01	0.005	30	7	3	5	0.02	0.015228426	0.235778342
Increase T to 360	0.99	0.01	0.005	360	7	3	5	0.02	0.015228426	0.019787789
Decrease C_1 to 6	0.99	0.01	0.005	180	6	3	5	0.02	0.015228426	0.033868584
Decrease C_1 to 1	0.99	0.01	0.005	180	1	3	5	0.02	0.015228426	0.006095037
Decrease C_2 to 2	0.99	0.01	0.005	180	7	2	5	0.02	0.015228426	0.039339538
Decrease C_2 to 1	0.99	0.01	0.005	180	7	1	5	0.02	0.015228426	0.039255781
Decrease C_3 to 4	0.99	0.01	0.005	180	7	3	4	0.02	0.015228426	0.039395939
Decrease C_3 to 1	0.99	0.01	0.005	180	7	3	1	0.02	0.015228426	0.039313875
Increase r_e to .025	0.99	0.01	0.005	180	7	3	5	0.025	0.020304569	0.039555133
Increase r_e to .05	0.99	0.01	0.005	180	7	3	5	0.05	0.045685279	0.040214326
Increase r_e to .10	0.99	0.01	0.005	180	7	3	5	0.1	0.096446701	0.041532713
Decrease r_e to .01	0.99	0.01	0.005	180	7	3	5	0.01	0.005076142	0.039159616
Decrease T to 90 and Decrease C_1 to 1	0.99	0.01	0.005	90	1	3	5	0.02	0.015228426	0.012037789

Figure 16. Results of Sensitivity Analysis on Morey's Accuracy Estimation Method

Method	Initial Measurement	Final Measurement	Impact	Change
Decision Analysis	Accuracy = .05	Accuracy = .99	+0.94	+1880%
	Availability = .72	Availability = .90	+0.18	+25%
	Loss = (3.16M)	Loss = (.289M)	+2.871M	+91%
Data Quality Attributes	% meeting threshold = 13.29	% meeting threshold = 89.1	+75.81%	+570%
Mathematical	% inaccurate records = 4.4	% inaccurate records = 1.2	+3.20%	+72%
Data Quality Calculus	Temporal-Effect = Intolerable	Temporal-Effect = Tolerable	+2 levels	+200%
	Accuracy = Medium	Accuracy = High	+1 level	+50%
	Reputation = High	Reputation = High	no change	0
	Ω = Temporal-Effect = Intolerable	Ω = Temporal-Effect = Tolerable	+2 levels	+200%

Figure 17. Results of Applying Data Quality Remedies to the AF C^4I Database

validated the appropriateness of its use for this environment. Even if the resulting improvements were in error by an order of magnitude, the fact remains that the DQEF provided the basis for realizing significant improvement in specific data quality assessments.

In addition, these assessments were based on the current model of the AF C^4I Database environment, allowing the incorporation of empirical measurements into the assessment methods. A second application of the DQEF could be exercised on the future AF C^4I Database environment, and then compared and refined as the current data life cycle evolves. This would serve to further validate and refine the DQEF.

3.3 Draw Conclusions

Conclusions were drawn in step five of the method and are presented here. First and foremost, a determination was made as to whether or not the Measures of Success were met.

The first two Measures of Success were binary decisions. It was shown that the DQEF allowed the definition and analysis activities to occur; therefore both of these Measures earned positive results. The third Measure of Success relied on quantitative and qualitative measures, where either the majority of the qualitative/quantitative attributes improved or the majority of the qualitative/quantitative attributes did not improve. Seventy five percent of the qualitative measures improved by at least one level after the implementation of remedial actions. All quantitative measures improved by at least 5% after the incorporation of remedial actions. The results indicate that all three Measures of Success were met or exceeded.

3.4 Make Recommendations

The major recommendation resulting from this research is to employ the DQEF to:

1. model the data environment, where the (current and future) context within which the data operate is described in terms of functions, users, information and data requirements, and business rules,
2. define the appropriate data quality attributes peculiar to the environment described in the model,
3. analyze the data environment using methods appropriate for the environment described in the model, and
4. improve data quality by incorporating suitable remedies.

The DQEF is flexible enough to accommodate general data-centric paradigms and many specific tools to attack data quality problems. By preceding the traditional "define, analyze, and improve" cycle with a modeling phase, the data quality engineer, as well as data users, become aware of the complete data environment. The suitability of specific methods for each phase of the DQEF becomes apparent much more readily once the environment has been defined. The fact that the DQEF considers temporal issues in the evolution of a data environment renders it a valuable tool in identifying and avoiding future data quality issues.

In addition, there are several areas which are beyond the scope of this research that warrant further investigation. For example, the temporal aspect of the data life cycle has only been touched upon in this research. The idea of engineering data quality into current and future data life cycle environments needs further investigation, especially during the transition period when the future system is not fully capable.

Experience with requirements translation and synthesis proved to be extremely useful in this research. Because many times users are not aware of their data quality requirements, having a basis from which to launch an initial investigation proved to be very important. Further work in this area could expand upon the standardization of data quality attribute definitions so that user requirements may be more easily identified.

Further research into some of the measurement methods appears to be warranted. For example, an expansion of the Decision Analysis method to include several data quality attributes may be indicated. In addition, automated tools to assist the decision-maker in these areas could also be developed. As another example, the mathematical estimation model could be refined based upon empirical data.

As a final recommendation, the whole realm of data quality could benefit from the development of automated tools which assist the non-software oriented user in modeling data environments, identifying requirements, measuring processes and products, and improving the data life cycle. This is an area rich with possibilities.

4. Summary

This paper has described an approach to developing a general Data Quality Engineering Framework (DQEF). The need for the DQEF was demonstrated and the steps involved in developing, applying, and proving the DQEF as a viable process were described.

We have shown that the DQEF is a general framework which can be tailored to the data quality environment at hand, and is flexible enough to be incorporated into any data quality program. The DQEF provides a framework for integrating specific data quality definition, analysis, and/or improvement solutions, whether offered by management theorists, computer scientists, mathematicians, or commercial industry. The key to the DQEF's success is the modeling phase, where the data quality engineer draws together the data producers and users, the data structure, and the data purpose. This allows the data quality engineer to become thoroughly familiar with the current and future data environment. Consequently, the integrated solution space is pruned long before any specific solutions are applied. The most noticeable result is saving — savings in time, savings in equipment, savings in software, savings in rework, as well as possibly savings in lives. Equally important is the DQEF's emphasis on temporal issues, enabling the data quality engineer (in concert with data producers and users, data administrators, data maintainers, etc.) to positively influence the evolution of the data and their environment. The focus on temporal issues results in the identification of opportunities for streamlining data life cycle processes early on, again resulting in measurable savings. The DQEF should be considered a crucial element in any data quality program.

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, 1995, pp. 1-3.

[4] Cardinal Business Media, "Restoring Credibility to MIS," *MIDRANGE Systems*, vol. 8, no. 23, 15 December 1995, pp. 25.

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

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

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

[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," PhD Thesis, Texas A&M University, 1989.

[10] S. E. Hansen and D. M. Meyen, "Horizon Link User's Guide Version 3.1B," The MITRE Corporation, McLean, VA MITRE Technical Report 96W0000047, July 1996.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] T. A. Larsen, "A Study of Minor Groove Binding and Lexitropsin Target DNA Sequences by X-Ray Crystallography," PhD Thesis, University of California, Los Angeles, 1990.

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

[18] T. D. Leblanc, "Measuring the Quality of Educational Data in NWFP Pakistan," PhD Thesis, Harvard University, 1995.

[19] D. M. Meyen and M. J. Willshire, "A Data Quality Engineering Process," Proceedings of the 1996 Conference on Information Quality, Cambridge, MA, October 1996, pp. 221-36.

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

[21] QDB, "QDB Solutions Products and Services," Personal Communication, QDB Solutions, Cambridge, MA, 22 January 1996.

[22] SRA, "Data Quality Engineering: Imposing Order on Chaos," Briefing Slides, SRA Corporation, Washington, DC, 1996.

[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] 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.

[25] Vality, "Product Brief: The Integrity Data Re-engineering Tool," Vality Technology, Boston, MA, July 1996.