DOD Guidelines on Data Quality Management

Defense Information Sysetms Agency
Center for Computer Systems Engineering
5600 Columbia Pike
Falls Church VA 22041

Phil Cykana, Alta Paul, Miranda Stern

Introduction

As organizations downsize, merge, and consolidate, automated information systems (AIS) which formerly did not communicate are increasingly required do to so to form an interoperable and shared data environment. Focused on achieving this environment, data quality issues often surface as important factors in facilitating and/or inhibiting system integration, data migration, and AIS interoperability.

As the Department of Defense (DOD) has downsized its information processing centers and emphasized the optimal use of selected migration systems, AIS functional proponents and system managers have come to the realization that different business uses of data impose different quality requirements and that data that was of acceptable quality for one system may not be so in another. In addition, data that was of sufficient accuracy and timeliness for local use may not be acceptable at another site. Costs of inaccurate or inadequate data can be steep. Problems with data quality can result in tangible and intangible damage ranging from loss of customer/user confidence to loss of life and mission.

Managing data quality in the DOD is essential to mission success. It ensures that quality data supports effective decision making and that data gets to the right person at the right time. In the Department, data quality management is composed of disciplines and procedures to ensure that data are meeting the quality characteristics required for uses in Command and Control (C2) systems, Procurement systems, Logistics systems, and the range of mission support applications that facilitate mission readiness, reliability, and effectiveness. In addition, improvement of data quality is lowering the costs of automated support to the DOD functional community by streamlining the exchange of technical and management information and making information systems easier to use.

DOD Total Data Quality Management (TDQM)

Data quality management in the DOD is focused on the same problems and issues that afflict the creation, management, and use of data in other organizations. As illustrated in Table 1, DOD data quality characteristics and conformance measures are similar to those used to measure data quality in any AIS. What is, perhaps, not similar is the size of potential data quality issues. In the Department of Defense, we have thousands of automated systems supporting users across the world. For example, in the DOD we have C2 systems supporting the Commanders in Chief (CINCs) in Europe and the Pacific, procurement systems supporting thousands of buyers and contract administrators, and hundreds of logistics systems that are used to requisition, stock, store, and issue equipment and materiel to soldiers, sailors, and airmen throughout the world. Across these systems, the Department has been involved in describing ways to improve data quality, to ensure that: (1) users (customers) of data are involved in improving data quality, (2) predetermined requirements for excellence are defined in terms of measurable data characteristics, and (3) data conforms to these requirements.

In the DOD, Total Data Quality Management (TDQM) is a process to support database migration, promote the use of data standards, and improvement of databases in conformance to business rules. The DOD TDQM approach borrows from other TQM methodologies in that it applies human resources and quantitative methods to improve products and/or services. The TDQM approach integrates functional management techniques, existing improvement efforts, and technical tools in a disciplined and focused way to create and sustain a culture that is committed to continuous improvement.

Data Quality Characteristics	Description	Example Metric
Accuracy	A quality of that which is free of error. A qualitative assessment of freedom from error, with a high assessment corresponding to a small error (FIPS Pub 11-3).	Percent of values that are correct when compared to the actual value. For example, M=Male when the subject is Male.
Completeness	Completeness is the degree to which values are present in the attributes that require them. (Data Quality Foundation)	Percent of data fields having values entered into them.
Consistency	Consistency is a measure of the degree to which a set of data satisfies a set of constraints. (Data Quality Management and Technology)	Percent of matching values across tables/files/records.
Timeliness	As a synonym for currency, timeliness represents the degree to which specified data values are up to date (Data Quality Management and Technology	Percent of data available within a specified threshold time frame (e.g., days, hours, minutes).
Uniqueness	The state of being the only one of its kind. Being without an equal or equivalent.	Percent of records having a unique primary key.
Validity	The quality of data that is founded on an adequate system of classification and is rigorous enough to compel acceptance. (DOD 8320.1-M).	Percent of data having values that fall within their respective domain of allowable values.

Table 1: DOD Core Set of Data Quality Requirements

Figure 1 illustrates the TDQM process as described within the Defense Information Systems Agency (DISA) for use across the Department. It begins with establishing the TDQM environment by building up management and infrastructure support and moves to the identification and definition of data quality projects. The selection of appropriate projects leads to the implementation activities. Importantly, the TDQM process also provides for the evaluation of the data quality management process. The idea is to review data quality goals/benefits and to improve the processes used to manage data quality.

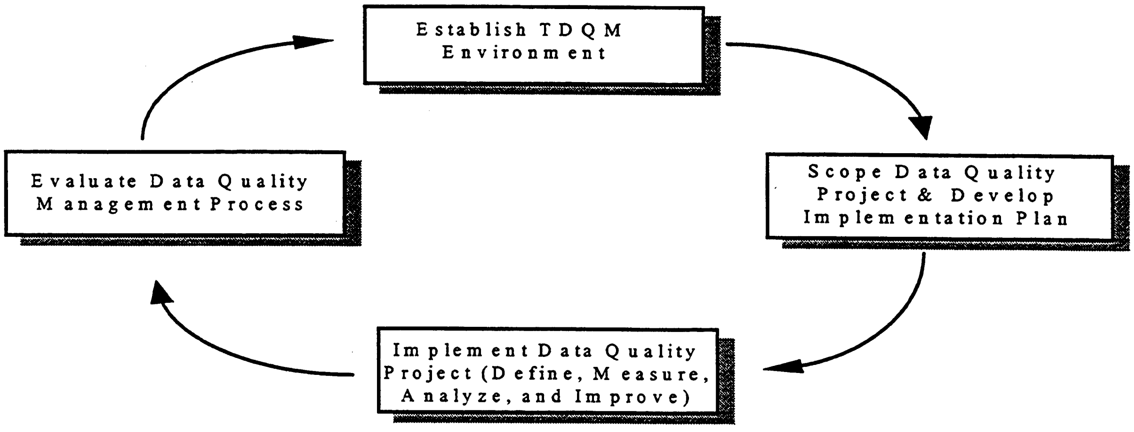
Establish TDQM Environment
Scope Data Quality Project & Develop Implementation Plan
Implement Data Quality Project (Define, Measure, Analyze, and Improve)
Evaluate Data Quality Management Process

Figure 1: DOD Total Data Quality Management Process

Establishing the TDQM Environment

As the first step in the TDQM process, establishing the data quality management environment is likely to be one of the most difficult steps in the process. Establishing the environment includes management buy-in and the cultural conditions that encourages team work between functional and AIS professionals. All to often, functional users of an AIS know the data quality problems that afflict an automated system but do not know how to systematically improve the data. In parallel, AIS professionals know how to identify data quality problems but do not know how to change the functional requirements that drive the systemic improvement of data. Given the existing barriers to communication, establishing the data quality environment involves the participation of both functional users and AIS administrators. In the DOD, this is accomplished by: (1) developing the strategic plan for data quality management and (2) developing and managing the cultural environment.

Developing the Strategic Plan

Data quality management responsibilities fall under the DOD information management and data administration initiatives (DODD 8000.1and DODD 8320.1). Key players in this initiative are the DOD Principal Staff Assistants (PSAs) (e.g., Acquisition and Technology (A&T); Command, Control, Communications and Intelligence (C3I); Directorate of Defense Procurement (DDP); Health Affairs (HA)) and their designated Functional Data Administrators (FDAds); Component Data Administrators (CDAds) (e.g., Army, Navy, Air Force, Marine Corps), and the CINCs (e.g., Atlantic Command (LANTCOM), Central Command (CENTCOM), European Command (EUCOM), Pacific Command (PACOM)). The FDAds, CDAds, and CINCs are responsible for developing data quality goals, objectives, and action plans for their respective organizations as part of their contribution to the DOD Data Administration Strategic Plan (DASP). The action plans are developed annually and provide information on:

- Overall goals and objectives for data quality management.
- Strategies and projects to achieve data quality goals and objectives.
- Measurable data quality objectives.

Developing and Managing the Cultural Environment

Action plans established by the FDAds, CDAds, and CINCs also address the infrastructure requirements to meet data quality objectives. Infrastructure needs include developing organizational responsibilities for improving data quality, establishing training programs and/or initiatives within functional areas, opening lines of communication between functional experts and AIS professionals about problems and solutions to poor data quality, and promoting functional and AIS improvements brought about by leadership to correct data quality problems.

Scoping the Data Quality Project and Developing the Implementation Plan

One of the major features of the TDQM approach that is promoted within DOD is the central focus on initiating and completing data quality projects. The essential requirement is to: (1) identify data quality improvement project(s) that can be successfully worked and (2) develop an implementation plan for each project.

Identify Data Quality Project(s)

Typically, data quality projects are selected by users and/or AIS administrators. It is good business practice to listen to both the functional and AIS community. For example, users often report frustration with errors in the data recorded in system tables and/or records. Known inaccuracies in queries, reports, and data correlation problems may be good indicators of data quality issues. Second, system administrators may make recommendations based on known problems with data collection, processing errors, and internal edit and validation procedures.

Additional factors that may influence the selection of data quality projects include focusing on areas that provide the greatest opportunity for success and prototyping/demonstrating the value of data quality efforts to achieve management buy-in.

- Choose Efforts that are Opportunities for Success: The success or failure of initial TDQM efforts or projects can greatly affect how easily the organization adopts TDQM ideas. Select projects: (1) that have a high chance of success, (2) that have the highest failure costs, and (3) where significant improvements can be made. Projects that address critical data quality issues that can be solved with the minimum of effort will increase the attractiveness of TDQM to top management.

- Prototype Effort: If there is not top management support for data quality efforts, perform a pilot project or demonstration. Choose a data quality project with low risk and low visibility that is critical to the organization's success. Focus on familiar data where functional and/or AIS expertise is readily available. Select an initial effort that is neither so large that it is doomed for failure from the start, nor so small that improvements will essentially go unnoticed.

Develop Data Quality Implementation Plan

Implementation plans are management documents that scope a data quality project in terms of project objectives, tasks, schedule, deliverables and resources. From a project management point of view, implementation plans provide information on:

- Task Summary: Lists project goals, scope and synopsis of anticipated benefits.
- Task Description: Describes data quality tasks.
- Project Approach: Summarizes tasks and tools to be used to baseline data quality.
- Schedule: Identifies task start, completion dates, and project milestones.
- Deliverables: Lists reports and/or products that document the result of a data quality project. At a minimum, deliverables should include:

  a. Data Quality Baseline Assessment - Document current data quality problems. Include exception reports on data that does not conform to established standards or business rules.

  b. After Action Report - Technical report on the data quality improvements that were implemented. Include description of actions taken to improve data quality and rationale for taking the actions and the lessons learned and improvement metrics.

- Resources: Identifies resources required to complete the data quality management project. Include costs connected with tools acquisition, labor hours (by labor category), training, travel, and other direct and indirect costs.

Implementing Data Quality Projects

Documenting the scope of a data quality project in terms of project objectives, tasks, schedule, deliverables, and resources provides the overall scheme for performing a data quality project. The execution of the project is the next step in the TDQM process. Generally, this step in the TDQM process is defined as consisting of the following four activities:

- Define: Identify data quality requirements and establish data quality metrics.
- Measure: Measure conformance with established business rules and develop exception reports.
- Analyze: Verify, validate, and assess the causes for poor data quality and analyze opportunities for improvement.
- Improve: Select data quality improvement opportunities that provide the most benefit and implement the selected improvements. Improving data quality may lead to changing data entry procedures, updating data validation rules, and/or use of DOD data standards to prescribe a uniform representation of data that is used throughout the DOD.

Generally, the four activities represent a robust set of activities that are designed to yield significant results in improving data quality.

Evaluating the Data Quality Management Process

The last step in the DOD TDQM process is the evaluation and assessment of progress made in implementing data quality initiatives and/or projects. Current DOD guidance encourages the participants in the TDQM process (FDAds, CDAs, CINCs, AIS functional proponents, and AIS administrators) to review progress with respect to: (1) modifying or rejuvenating existing methods to data quality management and/or (2) determining whether data quality projects have helped to achieve demonstrable goals and benefits.

The evaluation and assessment of data quality efforts reinforces the idea that TDQM is not a program but, a new way of doing business. In terms of evaluating and assessing progress made on data quality, the DOD FDAds, CDAds, and CINCs are encouraged to review both the costs and benefits associated with the data quality projects.

Executing the DOD Data Quality Management Methodology

The overall objectives of the DOD TDQM approach are to assess and validate data quality problems, identify root causes for data quality problems, and improve the quality and utility of data in DOD AIS. To meet these objectives, Figure 2 illustrates the four essential tasks connected to performing data quality work.

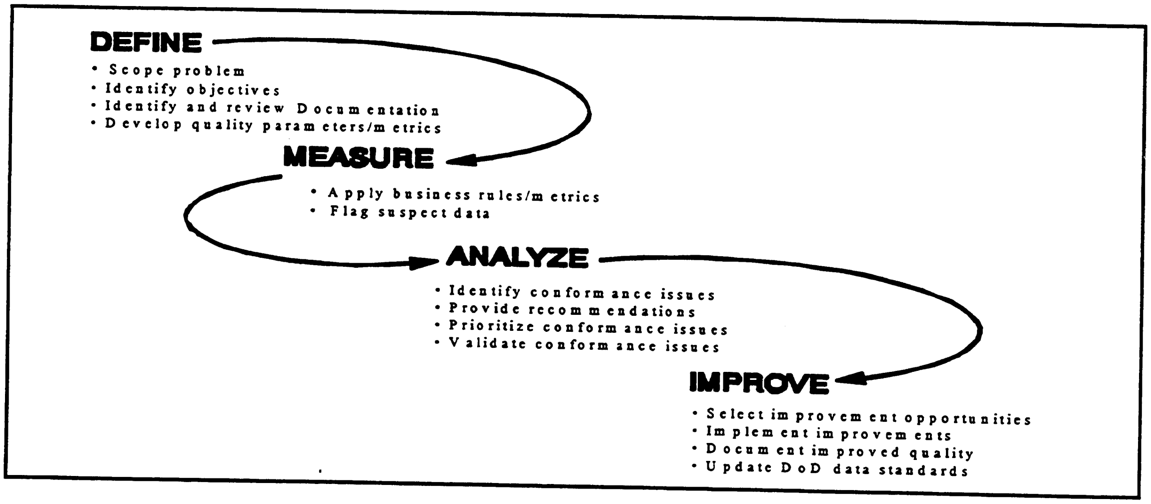
DEFINE
• Scope problem
• Identify objectives
• Identify and review Documentation
• Develop quality parameters/metrics
MEASURE
• Apply business rules/metrics
• Flag suspect data
ANALYZE
• Identify conformance issues
• Provide recommendations
• Prioritize conformance issues
• Validate conformance issues
IMPROVE
• Select improvement opportunities
• Implement improvements
• Document improved quality
• Update DoD data standards

Figure 2: DOD Project Implementation Guidance

As illustrated in Figure 2, DOD guidance is to: (1) DEFINE the data quality problem by establishing the scope of the data quality management project, objectives to be achieved by the project, and the criteria to be used to judge conformance to data quality standards; (2) MEASURE the conformance to data quality standards and flag exceptions to established data standards; (3) ANALYZE conformance and prioritize conformance issues to provide recommendations for improving data quality; (4) IMPROVE data quality by implementing recommendations.

Focusing on project execution, Figure 3 provides additional information on the stakeholders that are involved in improving data quality. Generally, the FDAd, CDAd, CINC, and/or the functional proponent for an AIS are pivotal players. These functional organizations typically provide sponsorship for DOD data quality efforts and are in the best position for identifying AIS data quality problems. Functional participation is also pivotal from the standpoint that sponsors mobilize functional subject matter experts that are used to support the improvement of data quality by specifying the business rules that are used to measure conformance to standards.

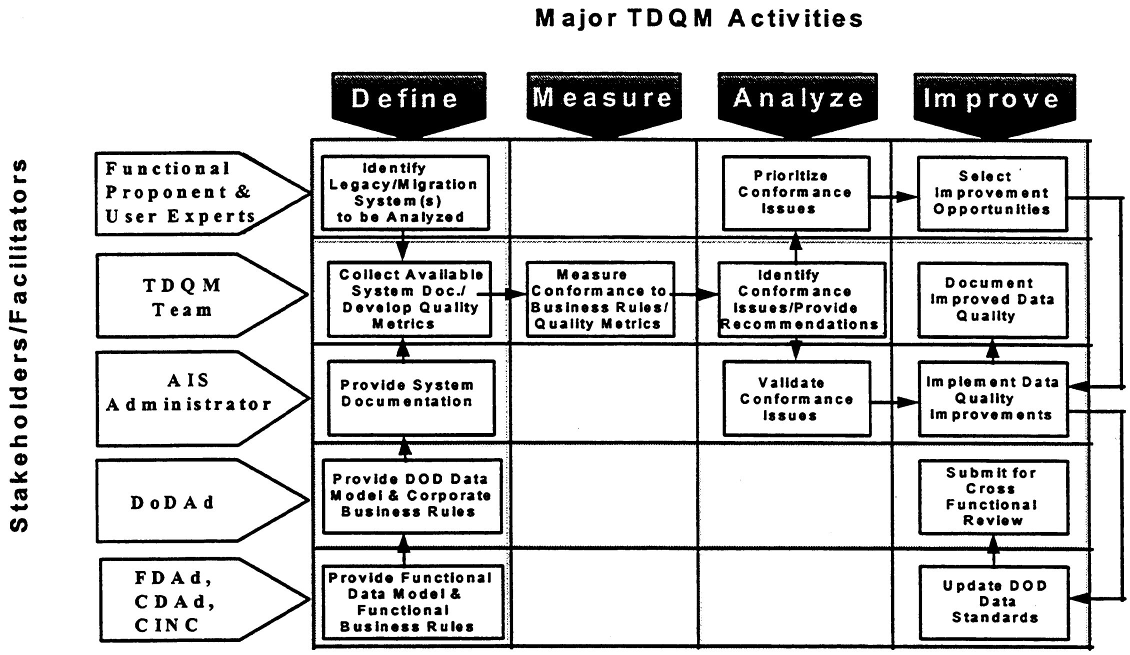
Major TDQM Activities
Define
Measure
Analyze
Improve
Stakeholders/Facilitators
Functional Proponent & User Experts
TDQM Team
AIS Administrator
DoDAd
FDAd, CDAd, CINC
Identify Legacy/Migration System(s) to be Analyzed
Collect Available System Doc./ Develop Quality Metrics
Provide System Documentation
Provide DOD Data Model & Corporate Business Rules
Provide Functional Data Model & Functional Business Rules
Measure Conformance to Business Rules/ Quality Metrics
Prioritize Conformance Issues
Identify Conformance Issues/Provide Recommendations
Validate Conformance Issues
Select Improvement Opportunities
Document Improved Data Quality
Implement Data Quality Improvements
Submit for Cross Functional Review
Update DOD Data Standards

Figure 3: DOD Data Quality Management - Stakeholders and Facilitators

In establishing measures of conformance, the DOD Data Administrator (DODAd) and AIS administrators also play a role. For example, within the DOD, the DODAd provides data standards that are used to measure conformance to approved standard data elements. In addition, AIS administrators provide important information on the business rules embedded in how data is managed by the AIS.

Define Data Quality

Defining the data quality for any given AIS is not a trivial task. The detailed description of specific data quality problems that are to be addressed by the project requires: (1) an analysis of historical data problems, (2) identifying and reviewing AIS documentation, and (3) the capture of business rules and data quality metrics. As a team effort, specific data problems are linked to business rules and both generic and specific rule sets are established to measure how good the data is within an AIS. Several rule sets are illustrated in Table 2.

Historical Data Problem	Rule Type	Generic Rule Set	Specific Rule Set
The equipment identifier fields are often blank.	Null Constraints	If the equipment identifier is zero, blank, or null then error.	Select equip_id from equip where equip_id = 0 or equip_id = ‘ ‘ or equip_id = NULL;
The interchangeability and substitutability (ISO) codes are not valid.	Domain Validation	If ISO code is not ‘B’, ‘I’, ‘G’ or ‘M’, then error.	Select iso_cd from equip where iso_cd not = ‘B’ or ‘I’ or ‘G’ or ‘M’;
The value of unit price is not greater than zero.	Operational Rule Set	If unit price = $00.00, then error.	Select * from equip where unit_price = 00.00;
The unit price for direct material is less than $10.00.	Relationship Validation	If material classification code equals ‘D’, then unit price must be greater than $10.00	Select * from equip where mat_class_cd = ‘D’ and unit_price < 10.00;

Table 2: Examples of Data Quality Rule Set Generation

DOD guidance on establishing rule sets encourages the development of data quality measures that can be executed in an AIS as actual code or as data quality filters in a data quality assessment tool. The rule sets that are developed represent the data quality metrics that are used to judge the conformance of data to the business rules. In the Department, data quality projects are also encouraged to make use of DOD data standards as the basis for establishing rule sets. For example, DOD data standards provide valid values for hundreds of data elements that are used in the Department. These include such domain sets as Country Code, US State Code, Treasury Agency Symbol Code, Security Level Code, Contract Type Code, and Blood Type Code.

Measure Data Quality.

The measurement of data quality in an AIS is to determine the exact nature and magnitude of data problems with the real data values stored in the tables/files/records that support an application. Measurement is the process by which the actual data instances are compared to the rule sets that were established as data quality metrics. Initial measurements of data quality are performed to establish the data quality baseline. Sampling techniques are encouraged to provide a valid baseline of data quality.

In terms of measuring data quality, there are two predominant approaches that are used in the DOD. The first approach is to measure conformance to data quality standards by executing the rule sets on the same machine and/or data server that supports the AIS. The performance of data quality checks are written as SQL scripts to test data conformance. The second approach to measuring conformance is shown in Figure 4 and is used in data migration situations where the data is actually moved to an interim environment prior to loading data to a target system. First, as illustrated in Figure 4, data must be either extracted from the source data systems and/or accessed to provide the

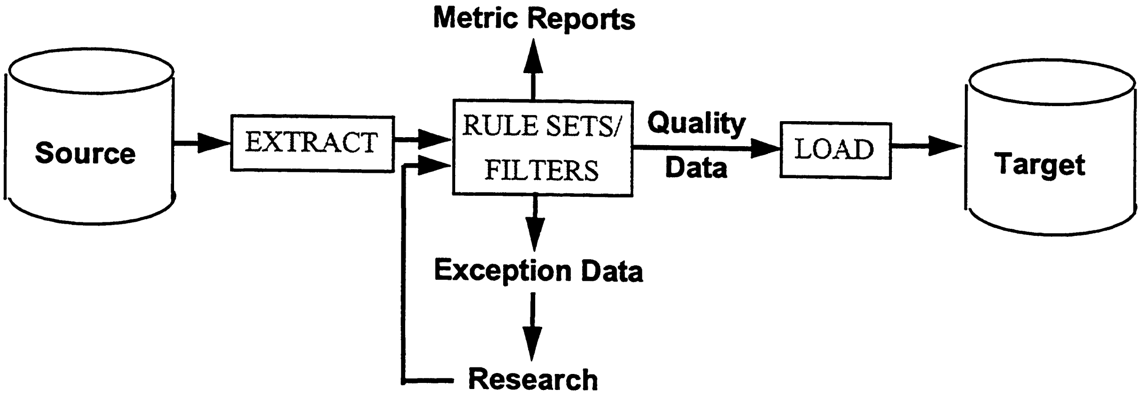

Figure 4: Performing Data Quality in Interim Environment

data sets that will be used. Second, the data sets are subjected to the rule sets or data quality filters that were developed to assess conformance to the established business rules. Third, exception data, or data that fails to pass the rule set is researched to determine why the data did not conform to the rule sets. In data migration situations, researched data are corrected and passed through the filter to check for errors. Also, one of the most important capabilities that is offered by this approach is the ability to generate metric reports. These reports may be used to provide statistical information on the conformance to data quality standards.

Generally, the measurement of data quality requires the performance of five activities:

- Determine the approach to be used to measure data quality.
- Apply the rule sets to the tables/files/records that are to be checked.
- Flag suspect data in error reports.
- Validate and refine the rule set.
- Develop metrics reports to categorize data quality problems.

Analyze Data Quality

The analysis of data quality problems relies heavily on metrics reports and the assistance of functional and technical data experts who are most familiar with the data and processes supported by an AIS. The analysis phase is devoted to identifying and validating: (1) key data quality problems from the metrics reports, (2) root causes for data quality problems, (3) cost impacts connected to correcting the root causes of data quality problems, and (4) solutions for improving the processes that are used to create and maintain data to minimize data errors.

Key Data Quality Problems

The analysis of metrics reports provides an opportunity to both identify and validate the types of data quality problems that exist in an AIS. As shown in Figure 5, metrics reports can provide an overall view of data quality within an AIS. Metrics reports also provide a method for measuring improvement which is based on the implementation of data quality process improvements.

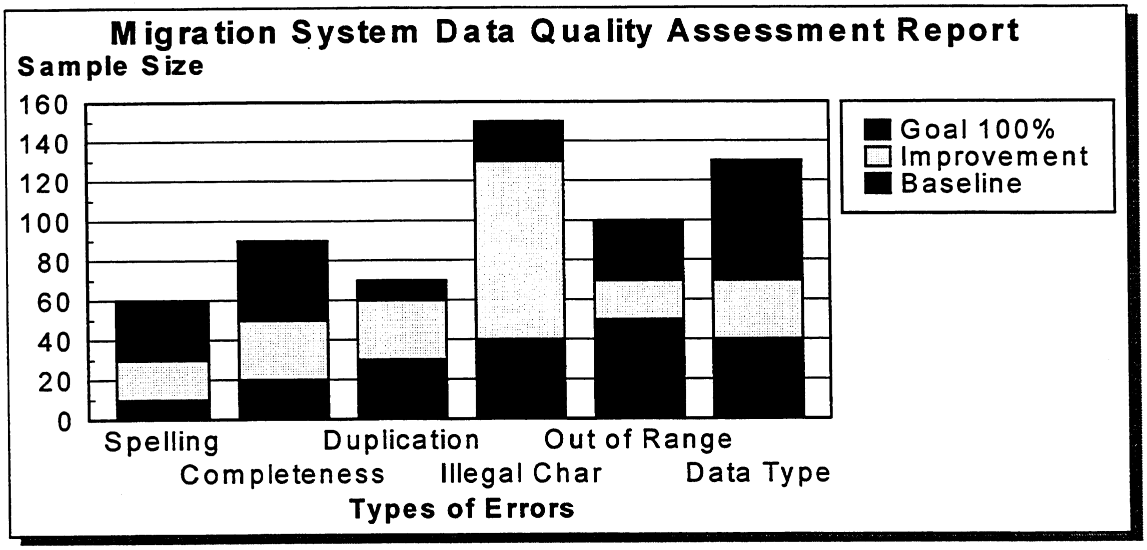

Figure 5: Sample Data Quality Metrics Report

DOD guidelines on data quality management encourage the use of data quality metrics reports to: (1) provide a baseline assessment of how good or bad the data quality really is within an AIS and (2) periodically check the data quality of an AIS to monitor progress towards attaining data quality goals. Graphical reports are recommended to provide a comparative basis of data quality trends and to compare reports to original baselines. Interestingly, our experience with data quality assessment tools tends to support their use over the development of SQL scripts and programming approaches to check data quality. For although scripts and programs can be written to execute data quality rule sets/filters, we have found that it is best to use tools that are specifically designed to perform data quality analyses with capabilities to: (1) audit the performance of data quality checks, (2) track historical records of prior data quality checks, and (3) graph data quality trends over time.

Root Causes for Data Quality Problems

The analysis of metrics reports and data quality trends provides an opportunity to assess the reoccurring problems that damage data quality. Some key questions that can be answered by the metrics information are:

- In what areas did a significant number of errors occur?
- Did certain types of errors occur more frequently than others?
- What is the best area on which to concentrate efforts so as to get the greatest improvement in data quality?

The analysis of errors that occur infrequently may reveal the cause of a specific error but, is not likely to identify a broad-based systemic problem. Nevertheless, fixing small problems (e.g., a one time data entry error) may offer anecdotal evidence that can be used to support the value of data quality assessments. The emphasis, however, of the DOD guidelines on data quality is to focus on root causes of data errors that are systemic. In determining these root causes, DOD guidance recommends the examination of possible causes for errors in data from several point of view. The four points of view include:

- Process Problem: Past experience has revealed that the majority of data errors can be attributed to process problems. For data errors categorized as process problems, DOD analysts are encouraged to examine the existing processes that support data entry, assignment and execution of data quality responsibilities, and methods used to exchange data. Knowledge of these activities in relation to data errors may be used to find and recommend actions to correct deficiencies.
- System Problem: Data problems often stem from system design deficiencies that are acerbated by poorly documented modifications and incomplete user training and/or user manuals, or systems that are being extended beyond their original intent. An examination of system modifications, user training, user manuals, and engineering change requests and problem reports can reveal AIS system problems that can aid in improving data quality.
- Policy and Procedure Problem: An analysis of data errors may reveal either conflicting guidance in current policy and procedure, lack of appropriate guidance, or a failure to comply with existing policy/procedure. An examination of existing directives, instructions, and standard operating procedures may be necessary to resolve the root cause of data errors.
- Data Design Problem: There is also the potential that the database itself will allow data errors to creep into data values as the result of batch loads, the use of incomplete data constraints, and/or the inappropriate specification of user privileges. An examination of batch load scripts or programs is recommended to eliminate possible data errors that can be attributed to circumventing data integrity constraints. It is also advisable to examine the implementation of: (1) primary key constraints; (2) null and not null data specifications, (3) unique key constraints and indexes; (4) database triggers; (5) stored functions and procedures; and (6) referential integrity specifications (e.g., cascading deletes). This technical information may reveal why persistent data errors compromise data quality in an AIS.

Assessing Cost Impacts

One of the real challenges in data quality management is the assessment of costs which are connected to correcting root causes for data quality problems and the costs associated with not correcting the problems that damage data. Measuring these costs is not easy. Guidance on collecting the costs connected to poor data quality focuses on defining the costs incurred to create and maintain the data and the cost of determining if the data values are acceptable, plus any cost incurred by the organization and the end user because the data did not meet requirements and/or end user expectations.

In general, the costs connected to poor data quality are categorized into four areas: prevention, appraisal, internal failure, and external failure. Once these costs are identified, there can be a better understanding of what it costs to correct data quality problems and what costs are incurred by ignoring data problems.

The main types of costs, depicted in Table 3, are direct and indirect costs associated with poor data quality. Direct costs include:

- Controllable costs: These are recurring costs for preventing, appraising, and correcting data errors.

- Resultant costs: Costs incurred as a result of poor data quality. Costs are considered internal failure costs and external failure costs.

- Equipment and training costs: Includes nonrecurring costs in data quality tools, ancillary hardware and software, and training required to prevent, appraise, and correct data quality.

In assessing direct costs, it is often useful to compare two or more alternatives for improving data quality and to estimate the controllable and equipment and training costs associated with each of the alternatives. This approach to costing corrective actions draws the attention of financial managers and accountants. It will require, however, work on estimating labor hours devoted to prevention, appraisal, and correction activities and estimates for equipment and training. The case study accompanying this document provides information on using direct cost estimates to estimate the costs connected to improving data quality.

DIRECT DATA QUALITY COSTS	INDIRECT DATA QUALITY COSTS
A. Controllable Costs	A. Customer incurred costs
1. Correction costs	
2. Appraisal costs	
3. Prevention costs	B. Customer dissatisfaction costs
B. Resultant Costs	
1. Internal-error costs	
2. External-error costs	
C. Equipment Costs	C. Loss of creditability costs

Table 3: Main Types of Data Quality Costs

Resultant costs and indirect costs are generally more difficult to quantify. Nevertheless, resultant and indirect costs should be used wherever possible to adequately assess the impacts of poor data quality. For example, the inability to match payroll records to the official employment record can cost millions in payroll overpayments to deserters, prisoners, and "ghost' soldiers. In addition, the inability to correlate purchase orders to invoices is a major problem in unmatched disbursements. In the DOD, resultant costs, such as payroll overpayments and unmatched disbursements, may be significant enough to warrant extensive changes in processes, systems, policy and procedure, and AIS data designs.

Recommending Solutions

The analysis of data quality is not complete until recommendations are provided on the actions to be taken to improve the data quality within an AIS. Recommendations should be supported by: (1) identifying the key data quality problems to be solved, (2) identifying the root causes for data quality problems, and (3) cost impacts connected to taking the corrective actions necessary to improve the data. Generally, if several alternatives are available, it is advisable to determine the level of risk that accompanies each alternative. Risk mitigation should favor small incremental improvements that are quick and easy to implement.

Improve Data Quality

Once recommendations have been made on the systematic actions to be taken to improve the data quality within an AIS, two additional major activities are performed. First, recommendations are usually reviewed by the functional proponent for the AIS and the AIS administrators to determine the feasibility of recommendations. The review of recommendations considers how solutions will affect end-users, functional processes, system administration, policy, and data design. Additional factors influencing the go ahead on recommendations include: (1) the availability of resources needed to accomplish the improvement, (2) the schedule of software releases, and (3) changes to the AIS hardware and/or telecommunications environment. Any one of these factors can influence the execution of data quality improvement recommendations.

The second major activity in improving data quality is to execute the recommendation(s) and monitor the implementation. In parallel with root causes for data quality problems, improvement work tends to fall into four categories:

- Process Improvement: Focus to improve the functional processes that are used to create, manage and use data. For example, functional process changes may encourage centralized data entry, elimination of nonvalue added activities, and the insertion of data quality responsibilities where data is entered into the AIS (e.g., certification of data)

- System Improvement: Software, hardware, and telecommunication changes can aid in improving data quality. For example, security software can be used to minimize the damage done by malicious updates to databases by unauthorized users. Hardware improvements may make batch loads faster and thereby make it unnecessary to turn off edit and validation constraints when loading data to a database. Telecommunications improvements (e.g., increasing bandwidth) may provide easier access to data and improve both the accuracy and timeliness of data. Other system improvements may include providing better end-user manuals, operation and maintenance manuals, and additional user training.

- Policy and Procedure Improvement: Resolve conflicts in existing policies and procedures and develop appropriate guidance that will institutionalize the behaviors that promote good data quality. One example is the development of Standard Operating Procedures (SOP) for an AIS that document the data quality rule sets/filters that are used to measure data quality. In addition, the performance of periodic data quality checks may be performed as part of the SOP to increase data quality.

- Data Design Improvement: Improve the overall data design and use DOD data standards. For example, database designs may be improved by the addition of primary key constraints, indexes, unique key constraints, triggers, stored functions and procedures, administration of user privileges, enforcement of security features, and referential integrity constraints. The use of DOD data standards supports the uniform representation of data across the DOD and supports improvements in data correlation.

Summary

DOD guidance on data quality management emphasizes the improvement of data quality to ensure that: (1) users of data are involved in the improving data quality, (2) predetermined requirements for excellence are defined in terms of measurable data characteristics, and (3) data conforms to these requirements.

The approach that has been adopted to achieve these goals consists of four steps. The first step is the establishment of the TDQM environment where key participants include the DOD PSAs, FDAds, CDAds, and CINCs. These key players are responsible for providing the overall direction for data quality initiatives and ensure that strategic plans and infrastructure elements are in place to support the improvement of data quality in the automated systems that support their functional mission. The second step in the approach is directly supported by AIS functional proponents and AIS administrators. These participants are responsible for identifying AIS data quality projects and the development of implementation plans. The third step is the meat-and-potatoes of data quality work. It consists of defining, measuring, analyzing, and improving data quality in selected automated systems on a project-by-project basis. Importantly, the emphasis of this step is to implement systemic solutions to data quality problems. Typically these consist of process, system, policy and procedure, and data design solutions that are tailored to institutionalize the conduct of a function to ensure the quality of data. The fourth step is an assessment activity that encourages the review of progress made with respect to: (1) modifying or rejuvenating existing methods to achieving data quality and/or (2) determining whether data quality projects have helped to achieve demonstrable goals and benefits.

Putting the TDQM approach to use within the Department, the DOD Services, Agencies, and CINCs have made important contributions to improving the quality and utility of data. In the future, data quality management will serve an increasingly important role in facilitating system integration, data migration, and AIS interoperability.

DATA QUALITY IMPROVEMENT CASE STUDY

1. INTRODUCTION TO THE PROBLEM

The Depot Maintenance Standard System (DMSS) is a DOD target migration information system that supports depot maintenance activities. DMSS replaces the functionality of multiple Air Force legacy systems at depot maintenance centers. The Depot Maintenance Management Information System (DMMIS) is a subset of DMSS and processes bills of materials (BOMs) and associated routing information for repair work orders as they progress through the work control centers.

The DMMIS data load project was initiated by the Joint Logistics System Center (JLSC) Warner-Robins Automated Systems Demonstration (WR-ASD) program manager at Wright-Patterson Air Force Base (WP-AFB). The overall DMMIS ASD project goal was to determine if automated methods and techniques could be developed to streamline loading data from legacy to target information systems. The Defense Information Systems Agency (DISA) provided data quality tools, guidance, and technical support during this project at the request of the WR-ASD program manager. The DMMIS data load addressed moving BOMs from legacy systems for the gyroscope repair facility at Warner-Robins Air Force Base (WR-AFB) into DMMIS. The BOM routing data for the work control centers will be addressed at a later time. The automated data load scenario replaces the difficult and manually intensive process developed in the gyroscope repair shop to collect, analyze, verify, and migrate "active" BOM data from the legacy systems. Figure 6 illustrates this manual process.

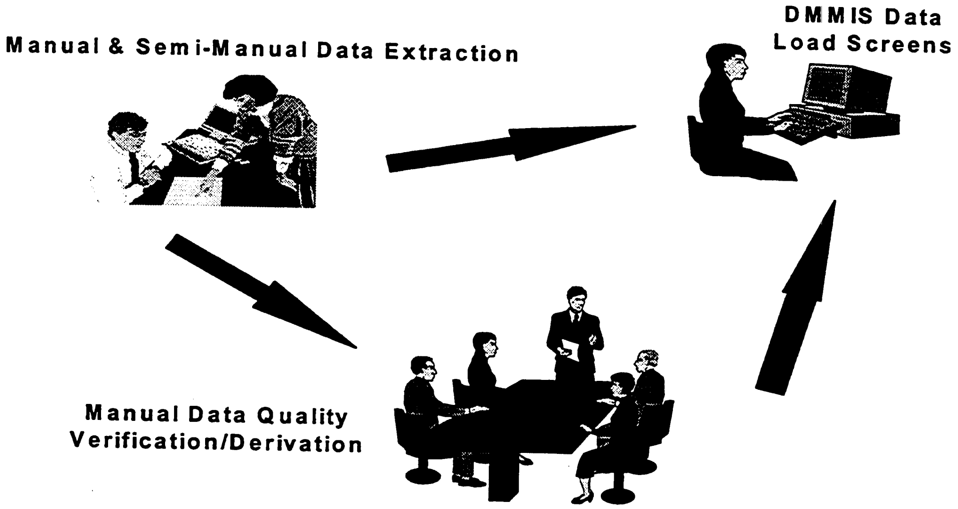
Manual & Semi-Manual Data Extraction
DMMIS Data Load Screens
Manual Data Quality Verification/Derivation

Figure 6: Manual DMMIS Data Loading Process

The DMMIS ASD goals were:

- minimize the need to manually re-key data from legacy systems into DMMIS,
- increase the efficiency of loading data into DMMIS, and
- improve and assure the quality of DMMIS data.

WR-AFB has 3500 "active" BOMs to migrate from legacy systems to DMMIS. Using the manual process, the JLSC projected that a significant cost in man-years and schedule would be needed to complete the DMMIS data load at WR-AFB. The cost estimate breakdown for this manual process was originally estimated to be:

- 3500 "active" BOMs with associated routing data at WR-AFB,
- 75-100 man-hours per BOM to manually load this data into DMMIS,
- 75 percent of effort is research, 25 percent data entry;
- Total estimate 350,000 man-hours (206 man-years), and
- Half of the work to load DMMIS involves migrating legacy BOMs.

The data quality project team applied an automated PC-based data quality analysis tool (QDB/Analyze ™) and DISA's Total Data Quality Management (TDQM) methodology to identify and address specific data quality issues in the legacy and target data environments that would negatively affect the efficiency and quality of the DMMIS data load effort.

2. THE DATA ENVIRONMENT

Access to legacy data during the DMMIS data load project was a major issue. The various stakeholders in the project were willing to provide the requested data, but the actual process to extract that data from the legacy environments proved to be very difficult. The major problem was obtaining supporting documentation on legacy system data structures and coordinating the resource requirements to perform the data extraction effort.

Data from two legacy depot maintenance systems was used during the project: G005M (Depot Maintenance Material Support System) and D043 (Master Item Identification System). The G005M data was extracted from a print file. The relevant data was parsed into relational tables which were then imported into the data quality analysis tool. The extracted G005M data also was used to retrieve D043 data for the component stock items on the BOM using the National Stock Number (NSN). A description of the legacy data used for data quality analysis in the project is shown in Figure 7

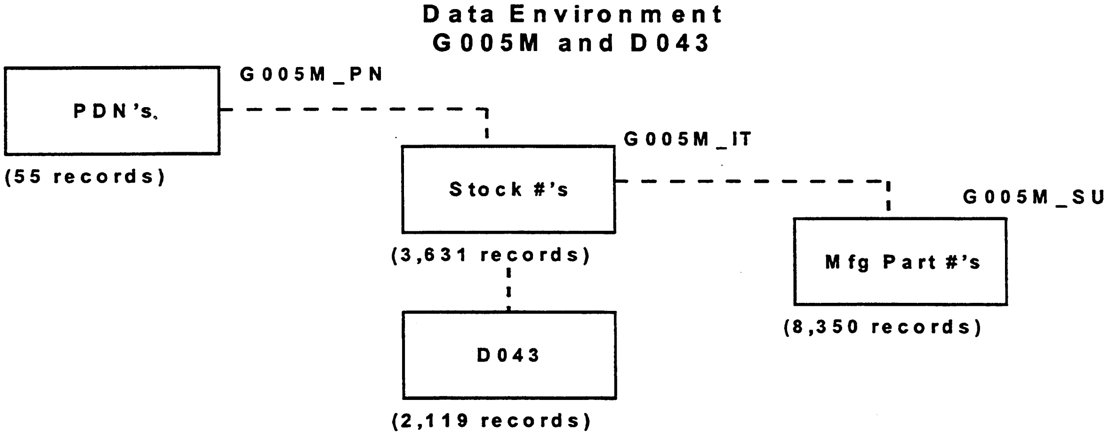

Figure 7: Relationships Among BOM Data

The DMMIS ASD project team selected 55 BOMs (designated by production number (PDN)) stored in the G005M system for this test. From these 55 BOMs, 3631 unique stock numbers and 8,350 manufacturing part numbers were found. Theoretically, each stock number should have had one corresponding record in the D043 system. After extracting the data, issues about the quality of the data from the legacy systems were examined.

3. LEVELS OF DATA QUALITY ANALYSIS

To achieve the goals of moving and improving the legacy data prior to migration, four levels of analysis were performed. The Level 1 analysis consisted of testing each data element from the sample set of completeness and validity checks. The completeness analysis determines for each data element the degree to which significant data values (i.e, not spaces, nulls, or defaults) are present in the legacy data. The validity analysis determines for each data element if those values are from the value set (domain) considered to be valid.

For the G005M and D043 legacy data analyzed, the completeness of data exceeded 99 percent for all data elements. The validity of the data elements varied from 80 percent to 100 percent. Samples of the data analysis are shown in Figure 8.

G005M_IT

	Data Field	Completeness	Validity
PDN	char(9)	100%	100%
END_ITM_ID	char(20)	100%	100%
OPER_NR	char(7)	100%	100%
ERC	char(2)	98%	98%
PSC	char(2)	99%	?*
FSCM	char(5)	100%	80%

*Documentation does not specify validity rules

Figure 8: Level 1 Analysis, Completeness and Validity

Level 2 analysis determines the structural and referential integrity of the data and defines the rules for cardinality found in and among the different data sets. The integrity analysis is used to validate the primary keys and their defining attributes (e.g., uniqueness) of record types and to validate the referential integrity of foreign keys within a record type. Analysis of cardinality determines whether records of a given key value must be unique in the table. If multiples of a given key value are allowed, a reasonable limit on the number of occurrences for the data set under analysis must be determined.

For the G005M and D043 legacy data, primary key integrity was good; however, referential integrity problems were prevalent across the G005M and D043 data tables. Some of these problems were later explainable by the staff at WR-AFB based on the way processing was being performed in the legacy systems during work control center repair operations. Cardinality analysis was not performed during the project. Figure 9 shows an example of a referential integrity problem identified. During level 2 analysis, the team discovered 103 NSNs in the D043 data that could not be linked to G005M records; and 67 NSNs in G005M were not found in D043 data.

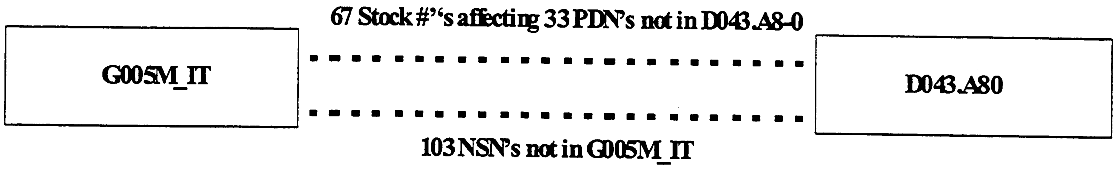
67 Stock #'"s affecting 33 PDN's not in D043.A8-0
G005M_IT
D043.A80
103 NSN's not in G005M_IT

Figure 9: Level 2 Analysis, Example of Referential Integrity Problems

Data models, data definitions, data base specifications, and other documentation are usually sufficient to complete most of level 1 and 2 analysis. Level 3 analysis requires direct support from the functional experts for the data set. For level 3, the rules of doing business in the functional experts' place of work are examined and turned into a series of automated tests. For this project, knowledgeable functional experts from the gyroscope repair shop at WR-AFB were interviewed. The team devised a set of rules about the data set under analysis that would effectively support the gyroscope repair shop planners in their analysis and verification of the legacy BOM data prior to the automated load process into DMMIS. The results from this analysis were not surprising since the project team zeroed in on known data problems in the legacy data in order to maximize the payback of the time invested for the study. Some of the business rules developed included valid combinations of values within data sets, computational verification within data sets, computational verification across data sets, and system flow point to point comparison.

Level 4 analysis involves developing the conversion rules for the legacy data and transforming the data into a format appropriate for the target system. A prototype system evolved from requirements and ideas identified by WR-AFB staff that resulted in a more efficient and effective data review process for planners preparing legacy data for DMMIS loading.

From the analysis, a methodology was developed to present the findings to the functional experts, the gyroscope repair shop planners. The prototype system processing consisted of the following steps:

- Select the BOM(s) to be reviewed,
- Analyze the BOM legacy data using a "filter" set of data rules defined during the four levels of data quality analysis,
- Build a table of BOM legacy data records which are appended with messages based on analysis, and
- Present the information to the planner to guide and facilitate preparing the DMMIS version of the BOM.

The WR-AFB staff reviewed the final design and operation of the prototype system and concluded that the principles of data quality as demonstrated in this project do indeed contribute value to the system solution of the DMMIS Data Load Automated System Demonstration.

4. PROJECT RESULTS

The project results successfully established the value of applying data quality analysis methods and techniques to the DMMIS data load process. Specifically:

- Problems were identified in the legacy data that would impact the integrity of data loaded into DMMIS,
- A feasible approach was developed to automate the planner's research and analysis of potential BOM data problems by developing and applying appropriate data "filters" in the data quality tool with the assistance of the subject matter experts at WR-AFB, and
- A high-level design was proposed for the DMMIS Data Load Automated Systems Demonstration solution that incorporates the use of automated data quality analysis as a key sub-system.

5. FINAL DMMIS DATA LOAD SOLUTION

At the completion of the prototype, the final design of the DMMIS data load solution and its integrated use of data quality analysis was developed. Data quality analysis will be performed as a sub-system function. Planners will not interface directly with the data quality tool but will be presented the data quality analysis results through workstation display screens. Figure 10 depicts how the technical solution brings together fragmented data sets from multiple sources into an interim database. The data quality subsystem provides an automated means to check data quality and prepare the BOM in DMMIS compatible format. The major benefit of the data quality subsystem is that it provides valuable information on what data from the legacy systems is problematic. Identifying questionable data allows the maintenance planner to focus efforts on researching problem areas and reduces the time and effort to prepare the BOM for DMMIS.

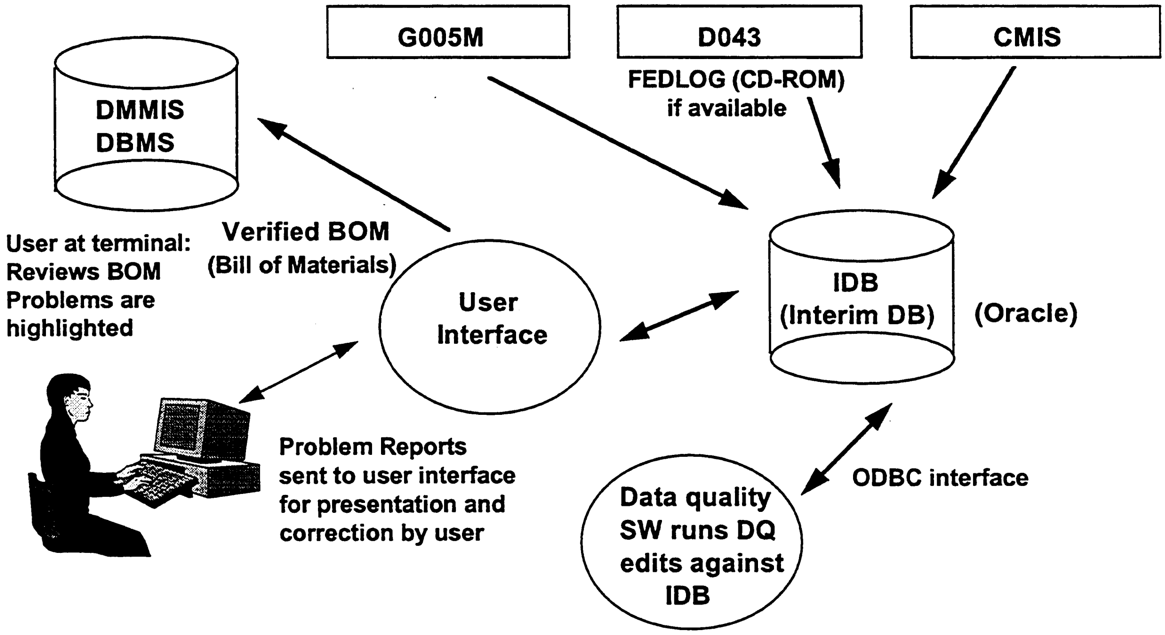
G005M
D043
CMIS
FEDLOG (CD-ROM) if available
DMMIS DBMS
Verified BOM (Bill of Materials)
User at terminal: Reviews BOM Problems are highlighted
User Interface
IDB (Interim DB)
(Oracle)
Problem Reports sent to user interface for presentation and correction by user
ODBC interface
Data quality SW runs DQ edits against IDB

Figure 10: DMMIS Data Load Solution

6. DATA QUALITY MANAGEMENT BENEFITS

Use of automated data quality analysis tools and TDQM techniques validated many of the principles and cost/benefits for performing data quality management. Some of the more quantifiable lessons learned and benefits defined by this project include:

- Rapid Results

A key feature of the project was the ability to quickly produce meaningful results that were easily understood by functional users and management.

- Cost Savings

Cost savings were projected in two major areas: (1) research costs devoted to identifying and fixing data quality problems prior to data load and (2) data entry costs devoted to manually rekeying data into the DMMIS. The initial projection was that research costs could be reduced by forty percent (40%) for the load of BOM data to DMMIS. In addition, data entry costs could be virtually eliminated. Table 4 shows projected costs broken out by existing data load methods and the recommended DMMIS data load solution. These projected savings were based on the original estimate of the work load to migrate 3500 "active" BOMs with relevant routing data into DMMIS.

Description/Cost Estimates	BOM Data Manual Research	BOM Data Manual Entry	Total BOM Data Load
Current Costs	$6.56M	$2.19M	$8.75M
Recommended DMMIS Data Load Solution			
Projected Reductions in:			
Time Saved	40%	100%	
Dollars Saved	$2.62M	$2.19M	$4.81M

Table 4: Projected Savings at Warner-Robins AFB

It was anticipated that some of these savings would be off set by both the nonrecurring costs for developing the DMMIS data load solution and by recurring technical and functional costs for operations and maintenance of the solution. Estimates on nonrecurring costs were placed at about $1.1 Million. Given these cost factors, net savings were projected at $3.71 Million at WR-AFB.

- User-Friendliness

The data quality tool chosen for this project was invaluable because of its functionality, cost, and user-friendliness. Of special note was the ability for end-user staff at WR-AFB to use the software for basic functions with minimal training. This experience validates that given training, end-users or systems personnel can be expected to carry out data quality analysis at the local site level where the data resides. This is key to the DISA TDQM approach which depends on a decentralized approach of data administrators, systems staff, and end-users to collaborate on pursuing data quality objectives on an ongoing basis. Specifically, for the DMMIS project, it also means that system maintenance of the data "filters" used in BOM analysis could be performed on-site by depot maintenance staff in lieu of making system change requests to remote system support organizations.

- Business Impacts

Using the automated data quality analysis tool and the structured levels of analysis enabled the project team to quantify the benefits of the project into a format for high-level discussions. This successfully elevates data quality to be one of the critical success factors in strategic systems technology efforts, and at the same time, educates senior management on the effects of data quality on the operations and mission of their organization.

7. CONCLUSION

Real savings of time and money were identified during this project by application of the techniques of TDQM. TDQM must become an active part of all data management projects. After test runs of the capability, the DMMIS Demonstration Results and Project Assessment Report provides the following:

"The use of the BOM Automated Data Load (ADL) tool provided better than a two to one improvement in productivity over manual methods during the migration of legacy BOM data into the target system, DMMIS. This, coupled with definite improvements in the BOM data quality observed by the team leaders, provides ample evidence that for large data load projects a significant reduction in migration costs would be achieved." (DMMIS Demonstration Results and Project Assessment Report, 3 June 1996)

Data quality techniques must be applied throughout the life cycle of information systems and then used to assist in migrating to the next generation of software that will support the same valuable assets - the data.

REFERENCES

Data Quality Foundation (Select Code No. 500-149), AT&T, 1992

DOD Directive 8000.1, *Defense Information Management (IM) Program*, April 14, 1992

DOD Directive 8320.1, *DOD Data Administration*, September 26, 1991

DOD 8320.1-M, *Data Administration Procedures*, March 1994

DOD 8320.1-M-1, *DOD Data Element Standardization Procedures*, January 1993

DOD *Data Administration Strategic Plan (DASP), Fiscal Years 1994-2000*, September 30, 1994

DMMIS Demonstration Results and Project Assessment Report, 3 June 1996

FIPS PUB 11-3, *American National Dictionary for Information Systems*, February 1991

Redman, Thomas C., *Data Quality Management and Technology*, Bantam Books, New York, 1992

Zero Defect Data Workbook: Conducting a Data Quality Baseline Audit, QDB Solutions, Inc., Cambridge, MA, 1991