AN INTRODUCTORY ANALYSIS OF THE HAZARDOUS MATERIALS INFORMATION RESOURCE SYSTEM: ESTABLISHING BASELINE MEASUREMENTS FOR DATA QUALITY AND APPLICATION OF THE TOTAL DATA QUALITY MANAGEMENT METHOD

(Research–in-Progress)

Mary E. Faber (Author)
Defense Logistics Information Services
Defense Logistics Agency
mary.faber@dla.mil

Kay Eggers (Briefer)
Defense Logistics Information Services
Defense Logistics Agency
Kay.eggers@dla.mil

Abstract: With the ever-changing field of information technology, organizations have greater capacity for data storage, manipulation, and sharing. Unfortunately, improvements have taken place so rapidly over the past decade that managers of database systems are realizing the growing need for data quality procedures. Many organizations have information systems that interface with other databases leading to issues of "dirty data" from upstream sources. Technology has allowed industry to share information, thus increasing operational efficiency, yet it contains errors and inconsistencies. Recent research has concluded that organizations must manage their data as a product, an integral part of supply chain management. The Total Data Quality Management (TDQM) methodology outlines systematic procedures for organizations to define, measure, analyze, and improve their data. This paper describes the data quality issues that the Defense Logistics Information Service is experiencing with the Hazardous Materials Information Resource System database. By utilizing the TDQM method, it also addresses baseline measurements, root causes, proposed solutions, and procedures for systematic measurement and analysis on an ongoing basis.

Key Words: Data Quality, Information Quality, TDQM

INTRODUCTION

As technology constantly evolves, it impacts existing information management systems. Improvements have allowed for information sharing through interfaces, data warehouses, and, most recently data enterprises. Unfortunately, much of the data may have originated using various older systems and over time the data may have become corrupt due to age, system errors, or poor quality edits from the original source. Data quality is a growing issue for all industries, and is no longer the sole responsibility of information technology (IT) departments. High profile initiatives such as Customer Relationship Management (CRM) and data warehousing have exacerbated the problem. An example of the repercussions is apparent in the case of a major electronic component manufacturer. Problems linking two internal databases led to unfulfilled customer orders resulting in a $2.5 million loss in revenue [1].

The Hazardous Materials Information Resource System (HMIRS) of the Department of Defense is no exception. HMIRS is the database and primary source of information regarding items of supply purchased by the government that contain hazardous material. It is designed to meet Occupational Safety and Health Administration (OSHA) standards as it contains Material Safety Data Sheets (MSDS) for every item as needed. In addition, value-added data is also entered for the items which include transportation, storage and other data that may be useful to the person in the field.

In response to Occupational Safety and Health Administration (OSHA) laws, the Department of Defense created HMIRS as a way to document and store Material Safety Data Sheets (MSDS) for items purchased by the government. HMIRS is a database in which data is entered both manually and automatically; therefore it is susceptible to errors. A data quality study is necessary in order to locate and assess areas of need as well as to understand potential consequences of low quality data. Such an analysis gives credence to decisions on tradeoffs between the cost of low quality data, the cost to improve data, and the return on investment. Before these can be measured, however, one must define data quality, establish a baseline, and determine the root causes.

Information has become a critical piece of the organizational structure. It is the foundation of mission statements, the driver behind strategic goals, the center of enterprise initiatives, and for some the basis of daily operations. The rapid pace of information technology (IT) has placed database management systems (DBMS) in the spotlight; however, some organizations are ill-prepared for the consequences. Gartner, Inc., a DBMS consulting firm, asserts that 25% of critical data within Fortune 1000 enterprises will continue to be inaccurate through 2007 [2]. In fact, a 2001 study found that poor data management is costing global businesses more than $1.4 billion per year as a result of faulty data in the areas of billing, accounting, and inventory [3].

It seems paradoxical that organizations that are so dependent on information have failed to give adequate attention to the quality of the data. Whether it is a lack of awareness, absence of processes and tools, or insufficient budget allocations, the cost of low quality data may have far reaching effects. It is important to realize that in addition to loss of revenue, organizations risk loss of opportunity, poor consumer perception, and sometimes even the risk of personal injury or lawsuits due to erroneous data. Such is the case with the Hazardous Materials Information Resource System (HMIRS).

BACKGROUND

DLIS Background

The Defense Logistics Information Service (DLIS) is an installation of the Defense Logistics Agency (DLA) under the Department of Defense (DOD). DLIS is housed in the Hart-Dole-Inouye Federal Center (HDI) in Battle Creek, Michigan. As the “information brokers” of the DOD it follows the DLA motto of “right item, right time, right place every time”. Its mission is “to create, obtain, manage, and integrate logistics data from a variety of sources for dissemination as user-friendly information to meet or exceed the needs of DOD, federal, and international logisticians” [4].

DLIS is comprised of seven Directorates, each having a unique role in the supply chain of supporting the military services and other DOD agencies. It manages the federal supply catalog system as well as many other databases that facilitate the supply, maintenance, and transportation of defense items. One such system is the Federal Logistics Information System (FLIS). It is a legacy system that provides automated data on the Federal Catalog System and descriptions of items of supply. It serves as a tool for purchase, stock, and storage of these items to facilitate logistics operations in the field.

The Armed Forces, federal agencies, NATO members as well as allies use the various DLIS managed systems. The rapidly evolving technology has changed the way DLIS conducts business over the last few decades, yet it continuously strives to be a premier player in the industry. Improvements have allowed for larger repositories, creation of data warehouses and partnering with other industry groups. DLIS Commander, Colonel Cassel, wrote "Major changes and improvements to the DOD supply chain continue to be underwritten by DLIS efforts in data interoperability, data and IT systems integration, and startling improvements in data quality" [5]. It is evident that data quality is an essential piece of the DLIS puzzle and appears to have support from the top down.

HMIRS Background

HMIRS was created in the mid 1970's and information was manually entered into a flat-file database. Very few system parameters existed, which allowed erroneous data to be keyed into the database. In May 2002, these files were migrated to an updated multi-dimensional Oracle-based system that allowed for greater automation and interfaces with other systems. As a result the "dirty data" that was stored in the old system was transferred and contaminated the new database. In addition, through automated processes, HMIRS began receiving secondary data from larger systems. Stakeholders realize that data quality issues exist in HMIRS; however, no formal measurement system had been created prior to this study. This paper outlines and analyzes a methodology that is suitable for measuring and tracking data quality. Using the Total Data Quality Method (TDQM) as described by Wang and Madnick, the paper defines, measures, and analyzes HMIRS data quality and then makes recommendations for improvement [6].

HMIRS data is entered by focal points from the four military services (Army, Navy, Marines, and Air Force) as well as General Services Agency (GSA) and the Defense Logistics Agency (DLA). A preliminary study to test the both the methodology and quality suggests that much of the erroneous data is being entered into the system by these focal points. Other root causes include outdated procedures and the transfer of poor data from other systems. The paper measures the error rate of ten data quality issues that fall into one of these three categories. The dimensions of accuracy, consistency, currency, and completeness are used to define data quality. The preliminary study included ten potential data quality indicators that were measured against these four dimensions.

The paper establishes a methodology for an empirical analysis of data quality. It provides tools to compare baseline measurements to benchmarks, and lastly categorizes issues to allow for systematic examination of root causes and potential solutions and concludes with an adaptation that is suitable for HMIRS. The Hazardous Materials Information Resource System is the database and source for complete product records for hazardous material used by the Department of Defense (DOD), General Services Agency, military services, and other federal agencies. Through the on-line website and CD-ROMs, it provides access to the Material Safety Data Sheet (MSDS) for hazardous items in the government inventory, as well as value-added data such as transportation, shipping, and storage information.

HMIRS includes a document submittal website that provides a publicly available location on the Internet where manufacturer, vendor, and government personnel can electronically submit MSDS and manufacturer's labels to HMIRS. It is at this point of entry by the focal points that erroneous data may be entering the system. The HMIRS database consists of over 330,000 records on items that are supplied by over 30,000 companies. Each record has up to 176 fields that contain varying amounts of data elements. HMIRS interfaces with five other systems, the largest being FLIS, which contains over a million records. Additional endeavors involving the inclusion of HMIRS in Enterprise Systems further exacerbates the data quality concerns. Due to the broad scope of HMIRS, it is evident that quality data is a necessity.

RATIONALE AND PURPOSE

Several issues and challenges currently affect the quality of HMIRS data. As previously stated, DLIS is not the data owner and, therefore, must work with the functional managers and authoritative sources to implement data quality initiatives. Some of the stakeholders have been resistant to change; therefore, it will be necessary to create "buy-in" of the data collectors and custodians.

HMIRS has evolved as a program since the mid 1970's, and rapid technological improvements have drastically changed the business practices. In the early stages few data limits were created and incorrect information was often entered into the database. Over time modifications were installed to prohibit incorrect data entry; therefore, newer data is much more accurate but the "dirty data" continues to exist. Although the various stakeholders have discussed quality issues and concerns, no formal study has been conducted. It is necessary to create a concrete methodology for establishing baselines in order to extract usable and valid results on the quality of HMIRS data. Such a study provides managers with a tool to accurately measure status and quantify improvement. It will provide a means to identify potential problem areas and where resources are best used.

Although DLIS is the program manager of HMIRS, it does not oversee the input of the data. As a result, data is often incomplete and/or incorrect. Vendors are given both Commercial and Government Entity (CAGE) codes and National Item Identification Numbers (NIINs) which interface with other systems. If these are incorrect, the result is missing or duplicate information in the HMIRS database and a discrepancy with interfacing databases. This can drastically affect the war fighter's ability to extract crucial information on handling hazardous materials in the field. The HMIRS team has identified potential areas where data quality may be compromised. This paper investigates the application of the Total Data Quality Management (TDQM) method for evaluating HMIRS data quality. TDQM includes a four-step process: *define, measure, analyze,* and *improve.* The objective is to construct a methodology for establishing baselines and benchmarks of current operations that can be used for further research of HMIRS data quality.

As a result of constant advances in the field of information technology, organizations are able to collect, manage, and share large volumes of data. Industry has responded with dynamic information systems (IS); however, it is only recently that the organizational missions and philosophies have begun to reflect the changes. In the past, total quality management (TQM) and an errorless system was the goal. TQM appeared to apply more to the manufacturing industry and created unrealistic expectations for data managers who do not have control over data sources. Research is revealing more innovative ways to approach dynamic information systems and the issue of data quality.

With the birth of IT and relational databases came the need for well-designed database management systems (DBMS). Codd's research of the Relational Model addresses the issue of quality in systems related data. [7] It was first created in 1969, and since then many derivations have emerged. It describes the relationship between domain values through the use of tables. Its function is to simplify data manipulation and facilitate data integrity for large databases. These relational tables were later used as the basis for commercially developed relational DBMS, such as Oracle and other SQL(structured query language)-based systems. It is here that the Total Data Quality Management model has its roots.

Total Data Quality Management (TDQM)

Organizations recognize that quality information is critical for success; however, most often it is treated as a by-product, rather than the product itself. In response to a lack of documented research, Wang and Madnick developed Total Data Quality Management (TDQM) as a method for measuring quality.[8] TDQM combines disciplines such as computer science, statistics, TQM, Codd's relational model, and organizational theory. It involves a four-step cyclical process that initially begins by defining data quality

parameters, establishing information quality metrics, analyzing root cause for problems, calculating their impacts, and lastly, identifying techniques for improvement.
The TDQM process crosses industry lines as it offers value for various organizations despite the differing missions [9]. It encourages tailoring the process specific to each organization's needs based upon two key steps. Organizations must clearly define "quality" in the general sense as it relates to their mission, and then apply specific data quality parameters. Secondly, it is important to develop a set of metrics that measure the important dimensions of data quality for the organization, and can be linked to the organization's general goals and objectives.

In his article, *A product perspective on total data quality management*, Wang focuses on delivering high-quality information products to consumers through a systematic and comprehensive data quality policy that is implemented top-down.[10] The TDQM method has roots in both TQM and the Deming cycle (plan, do, check, act). It consists of four phases of define, measure, analyze, and improve. In the first phase, *define*, data quality dimensions are established. The second phase of *measurement* includes identifying the baseline and information quality (IQ) metrics. This is followed by an analysis of the data and the root cause for data quality problems through statistical processes or trend analysis. Finally, techniques are implemented to improve the baseline measurements for better data quality according to organizational needs. This methodology has been researched over the last decade and successfully applied to various organizations.

Define

Quality has long been a focus of the manufacturing industry and is now becoming an integral part of computer-based systems. Although IT organizations service a variety of consumers, basic industry standards and definitions have crossed industry lines. Data quality is a multi-dimensional concept that varies according to perception and context. Pipino, Lee, and Wang address the concern of subjectivity by identifying and defining fourteen data quality dimensions.[11] The practice of defining what quality means to an organization is the first step of the TDQM process. As the TDQM tag line states, 'you can't measure what you can't define'. Organizations must begin by establishing quality definition as it relates to the organization's mission and vision. Creating operational definitions and parameters narrows the scope and ensures greater intra-organizational uniformity as it standardizes the focus of exactly what is to be measured.

Lee & Strong address the issue of defining data quality through multivariate dimensions.[12] Their research focused on five dimensions of data quality: accessibility, relevancy, timeliness, completeness, and accuracy. Next, they analyzed the relationship between knowledge of these data quality dimensions and various work roles within organizations. The researchers went beyond just analyzing data quality dimensions, but also incorporated how differing work roles view their importance relative to data quality. The authors define three work roles: *data collectors* (people, groups, and other sources that generate and input data), *data custodians* (those responsible for data storage, maintenance, and processing); and lastly *data consumers* (people or groups who use data)

The study found that the greater the knowledge of what, how, and why data were collected the better the quality. The knowledge held by data collectors played a key role in the quality of the data, while the greater custodian knowledge of the data production process led to higher degrees of accuracy, completeness, and timeliness. When comparing the three work roles, the authors found that the data collectors surprisingly had the largest affect on data quality, which contradicted previous assumptions. These findings are particularly important to HMIRS, as the various roles are divided in similar fashion and are located in different geographical locations. Many organizations focus too narrowly on accuracy alone.[13] It is necessary to have a much broader conceptualization that includes more than just one dimension of quality to adequately meet the needs of various consumers.

Measure

Data must be defined using a multi-dimensional approach, which include universal quality dimensions applicable across industries. The next step, *measure*, also has universal methods that can be tailored according to an organizations needs. There are three pervasive forms that have been found to be objective and widely adaptable: simple ratio, minimum or maximum operation, and weighted average. [16]

The simple ratio measures the number of undesirable outcomes (errors) divided by total outcomes subtracted from 1. The end product is a ratio depicting the number of positive outcomes. A calculation would be performed for each quality dimension. The second measurement method, *minimum or maximum operation*, is suitable for organizations using both multiple data quality indicators/variables and quality dimensions. The authors suggest computing the minimum or maximum values for each dimension by calculating the simple ratio of each dimension for the various indicators then comparing the values. The minimum calculation would be used for a conservative interpretation and the maximum for a liberal interpretation.

Lastly, the weighted average is another approach for the measurement of multivariate cases. Each dimension is given a weighted factor between 0 and 1 and will have a total of one. For example, if an organization has three dimensions and given the weights of .3, .5, and .2 respectively, the sum of these equals one. This method is useful when many quality dimensions are being used yet some have more importance and relevance than others relative to the indicators.

Analyze

The third step of TDQM is *analyze*. Upon establishing a baseline measurement, it is necessary to identify the cause of the inadequate data. This can be done through the creation of a data production flow map, also known as an information product map. Mapping data flow indicates where data is manipulated and if changes occur as a result. It allows managers to be more proactive as they identify where potential problems may arise. Furthermore, data maps assist in narrowing the scope as it defines and targets specific data elements. Networked environments are vulnerable; therefore, mapping is useful because it illustrates the inter-relationships of the systems due to migration, maintenance processes, and extracts.

Wand & Wang recommend establishing a model based on “ontological” concepts, such as using rigorous definitions of data quality dimensions by anchoring them in fundamental principles and production processes.[14] Upon establishing what quality means and how it is to be measured by using methods that are oriented toward system design, organizations must then begin the analysis. This involves identifying the problem, mapping the problem, assessing the reason for deficiency and finally making recommendations for data repair.

Improve

As the field of information technology has grown, organizations have failed to implement comprehensive methodologies for measuring data quality. For the past decade, most organizations have used ad hoc techniques that lack systematic measurement.[15] In the absence of adequate measurement procedures, organizations are unable to determine baselines, benchmarks, or assess progress. The growth of information systems has increased the need for high quality information, resulting in ever- growing IT budgets. Despite this fact, few organizations are measuring whether this money is well spent. Lee, Strong, Kahn, and Wang studied five organizations using a systematic methodology, which encompassed an information quality model, a questionnaire, and analysis techniques to assess information quality and benchmarks.[16]

In order to best analyze the gaps, the researchers used two techniques: benchmark gaps and role gaps. Benchmarking allows organizations to compare their performance against others in their field. It is defined as “a continuous systematic process for evaluating the products, services, and work processes of organizations that are recognized as representing best practices for the purposes of organizational improvement" (Spendolini as cited in Lee et al, 2002, p. 140).[17] Benchmark gaps assess an organization's data quality against an established benchmark (i.e., best practice). These are then plotted onto a graph to depict the size of the gap area, location of the gap (placement on the y-axis) and the different slopes over the x-axis. Below is an example of a benchmark gap graph. The researchers suggest using three indicators: the *size* of the gap area to determine whether various stakeholders are in agreement about data quality issues, the *location* of the gap on the quality scale, and the *direction* of the gap (positive versus negative).

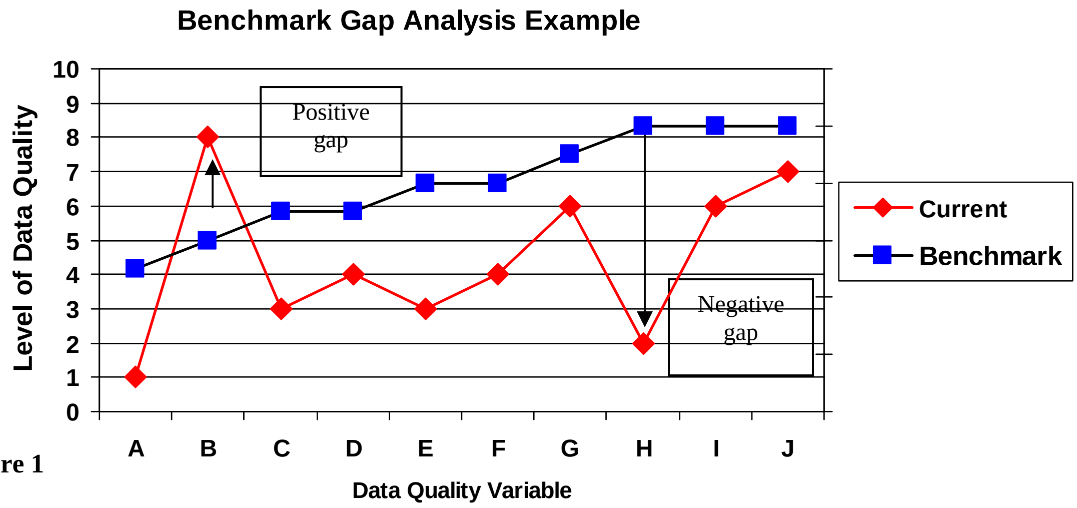


Figure 1

Studies have shown that a comprehensive approach is needed to address issues of data quality i.e., a methodology that goes beyond just accuracy and systematically addresses data quality issues using a multi-dimensional approach. Wang, Allen, Harris, and Madnick suggest using "total information awareness with quality" that is based on consumers' needs, managing information as a product, managing the life-cycle of the information product, and appointing information product managers. The multi-dimensional character of TDQM allows organizations to determine whether they are meeting the expectations of their consumers, how they compare with best practices, and accurately locate the problematic areas. Upon analyzing deficiencies, organizations can determine the root cause, and then finally concentrate their efforts towards improvement.

Application of TDQM

Data quality can be a very subjective and ambiguous area; however, it is possible to create objective metrics and baseline measurements by using the above approaches and standardized methodology. Additional research by Pipino et al appropriately validates the TDQM process.[18] The authors suggest performing both subjective and objective data quality assessments and comparing the results. Program managers must then identify discrepancies and determine the root causes. The last step is to identify and implement actions for improvement.

In their article, *A framework for analysis of data quality research*, Wang, Storey, and Firth provide a data quality framework that includes the following elements: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and a legal function.[19] The authors contend that these elements address the entire data life cycle from cradle to grave. It is important to note that this coincides with the DLA and DLIS philosophy of incorporating all aspects of the supply chain. In order to best capture the true measurement of an organization's data, each step in the data production process must be measured.

A useful example of successful implementation of TDQM is the SC Johnson Wax Company. In 1998, S.C. Johnson Wax (SCJ) revamped their business drivers and revealed the need for clearer goals and instituting TDQM initiatives.[20] The SCJ goal was to operate on a global base while reducing information management costs. They found that TDQM initiatives were necessary in order to improve customer, product, and vendor information. Just as with the HMIRS system, SCJ systems received data from external, internal and third party sources. Upon investigating current operations, SCJ found that approximately 60% of project time was spent on data cleansing due to information quality problems.

SCJ transformed their ad-hoc data quality projects into long-term systematic assessments using the four stages of the TDQM cycle. In response to their findings, SCJ incorporated two key tasks as part of their TDQM effort, an information quality audit and assessment. The first was an automated process that included using software to extract files and identify erroneous data in the system. The second task consisted of a training process for functional personnel who either enter or extract information from the system. The intent was to assess the perceptions and compare this to the "reality" of the audit information.

Although TDQM is an ongoing process, the S.C. Johnson Wax case study demonstrates a useful model for delivering quality information to internal and external customers. Although SCJ is a provider of household products, their issues are similar to those found in HMIRS. Incorporating TDQM methods in a similar fashion will perhaps have comparable results.

METHODOLOGY

A joint decision by management and the HMIRS program management office produced thirty-one data quality issues. These issues were chosen and prioritized based on linkages to the organizational vision, management theories, Federal Logistics Information Service (FLIS) interface, level of importance, potential hazard due to inaction, amount of DLIS control, and magnitude of the data quality issue.

TDQM Step1 - Define

The four dimensions of accuracy, consistency, currency, and completeness were chosen and defined by the Data Quality team.[21] This is different from TDQM as Wang and Madnick suggest using fourteen data quality dimensions. Upon review, however, the DQP managers decided that not all fourteen were applicable to DLIS systems, and the large number would impede implementation.

Accuracy - The measure or degree of agreement between a value (or set of values) and reality. The data is correct for what is being represented.

Consistency – The data passes all edits for acceptability. i.e., format, length, characteristics, values.

Currency – The data is up-to-date and the age of the data is appropriate for the task at hand.

Completeness – The measured data that should have values in them, in fact do so. Input would be based on customer/system needs [22].

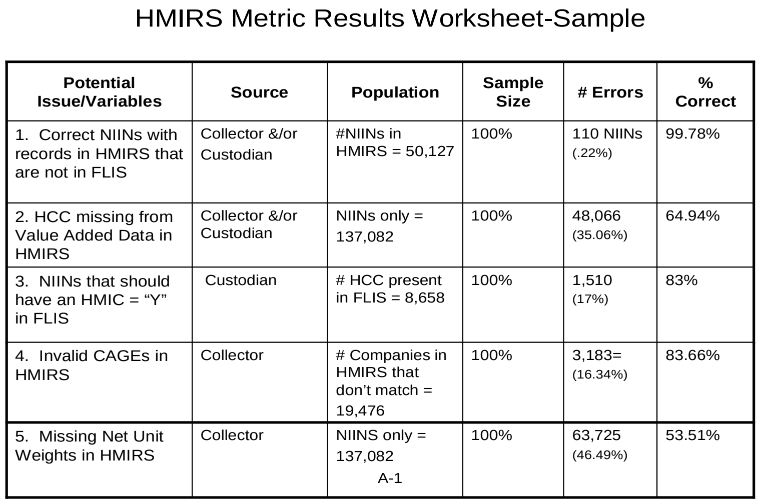
These dimensions were then applied to the thirty-one data quality indicators (variables), which are affiliated with the target population of the HMIRS database and its interface with FLIS. The thirty-one issues were then prioritized and ten of these were chosen as the variables for the study.

TDQM Step2 – Measure

Reports were extracted from the HMIRS and FLIS databases using Oracle-based software, SQL commands, and queries on data elements of the ten variables. The data extracts indicated the number of errors in the HMIRS database and the inconsistencies between the two systems. The number of errors was then divided by the population size to calculate the error percentage. The inverse (percent correct) was calculated by subtracting this number from one. Each variable was measured using the four dimensions. For example, calculations were performed to determine the error percentage for accuracy, consistency, currency, and completeness. The result was four error percentages for each quality issue. This was necessary as it is possible for a data element to be valid in one dimension but not the other (i.e., complete, but not current).

In order to account for the multivariate nature of the issues, weights were assigned to each of the four dimensions using decimal values for each dimension. The dimension error rates were then multiplied by the weight. The products of each calculation were then added for a sum total, which signified the percentage for that variable. This step was completed for each of the ten variables. When completed, however, it became apparent that the assignment of weights was rather arbitrary and subjective, thus vulnerable to researcher bias. As a result, an alternate method was used to account for the multiple variables and definitions.

Upon reviewing previous research, it was decided that the *Minimum Operation method* as described by Pipino et al, allowed for the most accurate yet conservative reflection of the quality of the HMIRS database [23]. After each of the ten indicators was measured separately according to the four dimensions, the error percentage was converted to percent of correct data by subtracting the error percentage from one.



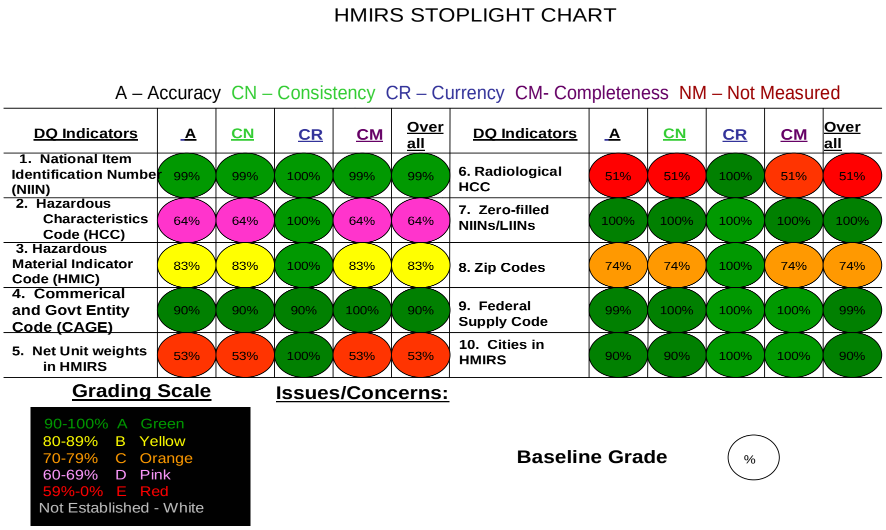
HMIRS Metric Results Worksheet-Sample

Potential Issue/Variables	Source	Population	Sample Size	# Errors	% Correct
1. Correct NIINs with records in HMIRS that are not in FLIS	Collector &/or Custodian	#NIINs in HMIRS = 50,127	100%	110 NIINs (.22%)	99.78%
2. HCC missing from Value Added Data in HMIRS	Collector &/or Custodian	NIINs only = 137,082	100%	48,066 (35.06%)	64.94%
3. NIINs that should have an HMIC = "Y" in FLIS	Custodian	# HCC present in FLIS = 8,658	100%	1,510 (17%)	83%
4. Invalid CAGEs in HMIRS	Collector	# Companies in HMIRS that don't match = 19,476	100%	3,183= (16.34%)	83.66%
5. Missing Net Unit Weights in HMIRS	Collector	NIINS only = 137,082 A-1	100%	63,725 (46.49%)	53.51%

Figure 2

Figure 2 is the template used to outline the thought process and identify the population and sample sizes used in the calculations. The aggregate percentages of each quality issue were charted and assigned a color-coded grade as shown in Figure 3. The HMIRS system can also be assigned an "overall" grade using this methodology; yet specific data are still visible, thus not lost in the process of "rolling up" the data. This illustration not only assists in the analysis of the data, but also allows for program managers and administration to quickly identify areas of success and those that require attention. Following the data collection and baseline measurements, benchmarks were also established. As the grading scale illustrates, a score of 90-100% represents the upper echelon, and is the benchmark for the data quality issues. It is realized that a goal of "perfect" data is not realistic. It is very rare that data is 100% correct. The benchmark for the ten HMIRS variables, therefore, is to score between 90-100%, thus receiving a grade of "A". The data issue then received the percentage of the lowest scoring dimension.

As illustrated in the chart below, the first variable received a score of 99%, as that was the lowest percentage for that issue. On the other hand the fifth data issue received a low score of 53%, which then equals the overall score. This method reflects the philosophy of an organization only being as strong as its weakest link. These calculations reflect the baseline metrics for the first ten data elements. One of the criteria for selecting the ten elements included the feasibility of extracting the data. The selected data issues were established elements in the HMIRS relational tables.



HMIRS STOPLIGHT CHART

A – Accuracy CN – Consistency CR – Currency CM- Completeness NM – Not Measured

DQ Indicators	A	CN	CR	CM	Over all	DQ Indicators	A	CN	CR	CM	Over all
1. National Item Identification Number (NIIN)	99%	99%	100%	99%	99%	6. Radiological HCC	51%	51%	100%	51%	51%
2. Hazardous Characteristics Code (HCC)	64%	64%	100%	64%	64%	7. Zero-filled NIINs/LIINs	100%	100%	100%	100%	100%
3. Hazardous Material Indicator Code (HMIC)	83%	83%	100%	83%	83%	8. Zip Codes	74%	74%	100%	74%	74%
4. Commerical and Govt Entity Code (CAGE)	90%	90%	90%	100%	90%	9. Federal Supply Code	99%	100%	100%	100%	99%
5. Net Unit weights in HMIRS	53%	53%	100%	53%	53%	10. Cities in HMIRS	90%	90%	100%	100%	90%

Grading Scale

90-100%	A	Green
80-89%	B	Yellow
70-79%	C	Orange
60-69%	D	Pink
59%-0%	E	Red
Not Established - White		

Issues/Concerns:

Baseline Grade %

Figure 3

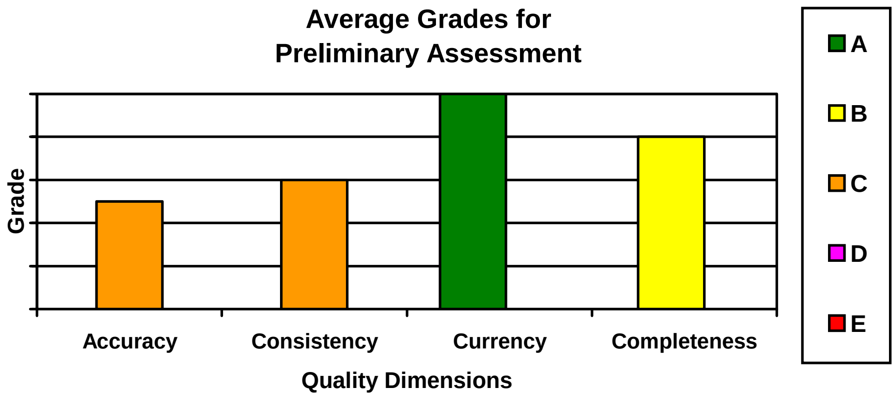
Data findings and the benchmark gaps will be measured on a quarterly basis in order to document progress or regression toward specified goals on a continual and regular basis. The data will continue to be collected using a time series design. The intervals will be charted overtime using a line graph depicting status of the ten data issues as they compare to the benchmark. Collecting data over time will allow the establishment of a trend line. Analysis will then portray whether or not implemented changes are meeting established goals and objectives.

RESULTS

Prior to the study, management believed that HMIRS contained large amounts of dirty data. According to the charts, however, the inaccuracies exist only in certain areas. The lowest scoring data issue is #5. Although it is 100% *current*, the other dimensions have substandard rankings. Of the ten variables, two received a grade of "E" (#5, #6); one received a grade of "D" (#2); one received a grade of "C" (#8), one received a grade of "B" (#3), and the remaining five data quality issues earned a grade of "A" (#1, #4, #7, #9, #10) as illustrated in Figure 3. Using these results, it is possible to calculate an average by adding the percentages and dividing by the number of data quality issues, which is ten. The product of the equation is 80.3%; therefore the overall grade for HMIRS using these ten issues is a "B".

It is also possible to assign "grades" to the four data quality dimensions using the same formulas. The percentages for each of the dimensions are added together and then divided by the total number of issues (ten) to determine an average as shown in Figure 4. The lowest scoring dimensions for the ten issues appears to be *Accuracy*, *Consistency* and *Completeness* with an average of 80.3%, which results in a grade of "B", whereas *Currency* received an "A", scoring 100%.

Figure 4



In order to make the data more practical, it is possible to categorize each of the ten data issues into three categories. This facilitates the analysis process and subsequent recommendations. Issues #9 and #10 are system related problems. Issues #2, #3, and #5 are process related issues and items #1, #4, #6, #7, #8, are caused by user/human error. The process of categorizing the data issues provides focus and allows one to hypothesize of the potential root cause. Figure 5 represents the measurement of the data quality issues classified by potential source.

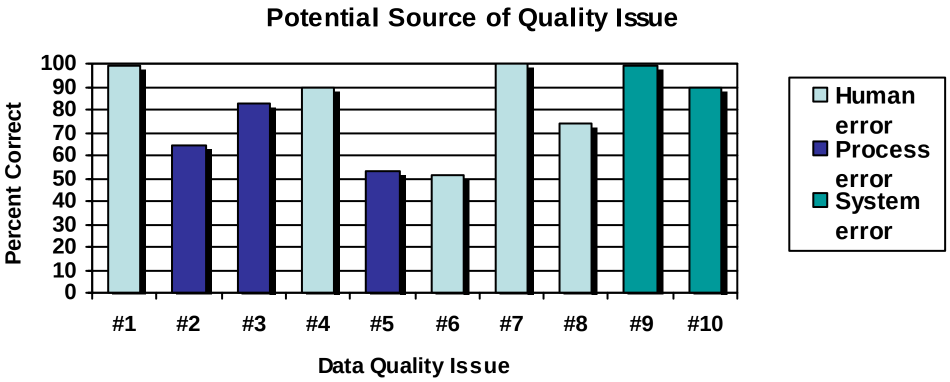


Figure 5

These categories coincide with research of the "three C's" of data quality: collectors, custodians, and consumers. Systemic issues are the responsibility of the custodian. Process related issues can be attributed to both collectors and custodians, while the data entry errors are the cause of the data collectors [24]. The figures represent the baseline measurements for HMIRS. Charting the data findings assists in performing a gap analysis between the baseline and the benchmarks. It also facilitates the analysis process by providing a starting point for determining the root cause of data quality problems.

TDQM Step 3 - Analyze

The third stage in the TDQM process is analyzing the data quality issues. In order to best analyze the baseline measurements, it is helpful to compare them to the established benchmarks. The graph below illustrates how HMIRS performed in the ten areas against the benchmark of scoring between 90-100% across all ten issues in each of the four dimensions.

It is apparent that a gap exists between current status and the benchmark. The graph illustrates the degree of variance between actual and desired quality. Such a graph is useful when measuring over time to chart progress or regression in the various areas. A noticeable drop or increase can be cyclical or seasonal; therefore, it is necessary to perform these measurements on a continual basis for a clear and accurate depiction of the system.

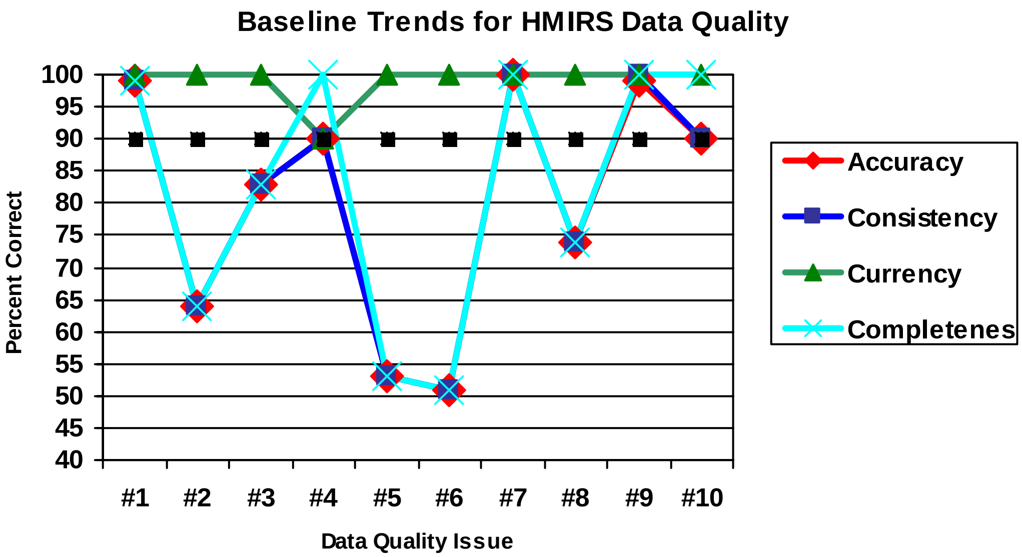


Figure 6

The findings show that *Accuracy* is clearly falling short of expectations; however, the analysis does not end here. It is necessary to trace the data element through its life cycle from cradle to grave in order to find the root cause of the inaccuracies and inconsistencies. Using the TDQM methodology, data production maps were created for each of the data elements involved in the ten data issues.

HMIRS Data Quality Indicators

DQ Indicator	Root Cause	Current Status	Proposed Solution
1. Correct NIINs with records in HMIRS that are not in FLIS	•Reference number (Part number and CAGE) are incorrect in HMIRS and therefore the information does not get transferred to FLIS during the interface •Focal point keyed in invalid NIIN •Invalid NIINs in HMIRS exist due to inaccuracies in the old HMIS system. These were migrated into HMIRS. •Invalid NIINs continue to exist because the info cannot be transferred/updated due to inconsistencies between the two systems •Discontinuing/canceling a NIIN in FLIS, yet the information must be kept in HMIRS for 40 years	•Discussed issue at HMIRS monthly IPR with FMO and contractors	•Identify what NIINs have been completely deleted from FLIS. •Weekly meetings with FLIS PMs •Review the original requirement for the data extract and update with SCR to prevent data from being overwritten in HMIRS •Research the DRN to find where HMIRS is pulling this data from FLIS

Figure 7

Following the analysis of these charts, the root causes were documented using the template shown in Figure 7. The cause for 40% of the issues was found at the point at which the data is entered into the system. These errors were due to the focal point keying in invalid data or in some cases failing to enter the data. This FLIS/HMIRS interface was the location of 30% of the errors, whereas outdated procedures in the functional process accounted for 20%. Lastly, 10% of the errors were caused by dirty data being migrated from the old HMIS system to the updated HMIRS.

As with many information management systems, upgrades are being developed on an ongoing basis. To account for this, the column "Current Status" was added to describe any system changes to upcoming releases that may correct the data issue. Although the data issue may be prevented in future releases, the dirty data still exists and, therefore, must be "cleaned up". As the chart illustrates, some of the solutions are proactive while others are reactive.

DISCUSSION

TDQM Step 4 - Improve

Data quality is a very comprehensive issue that necessitates ongoing measurement and assessment. The ten quality issues described in this paper are not an exhaustive list. Therefore, it is necessary to solicit feedback from consumers to determine their perceptions of data quality problems with HMIRS. Conducting a consumer survey would facilitate this process; however, this is a very time consuming process. The process of creating a survey should be done carefully in order for the results to be useful. A consumer survey would allow the HMIRS managers to compare the users' data quality perception versus reality. Differences should be charted using a role gap analysis. Survey results would also reveal consumer requirements, expectations, and wants. This would facilitate the actions of other departments. If results illustrate average data quality but poor consumer perception, a marketing plan that highlights successes and dispels myths would be very beneficial. Comparing future survey results to data quality research would assist in budgeting for future projects and endeavors. It may be possible that data quality is sub-standard in a particular area; however, it is an issue that is given low priority by stakeholders. Resources, both funding and personnel, may be better spent and utilized in other areas that have greater value to all parties involved.

Due to time constraints however, the PMO instead solicited input from the various data collectors and custodians. Some may feel this step should have preceded the data quality study. On the contrary, the intention of this project was to serve as a preliminary study to provide insight as to whether further research was needed, and if so, to determine the key areas that needed more detailed inspection.

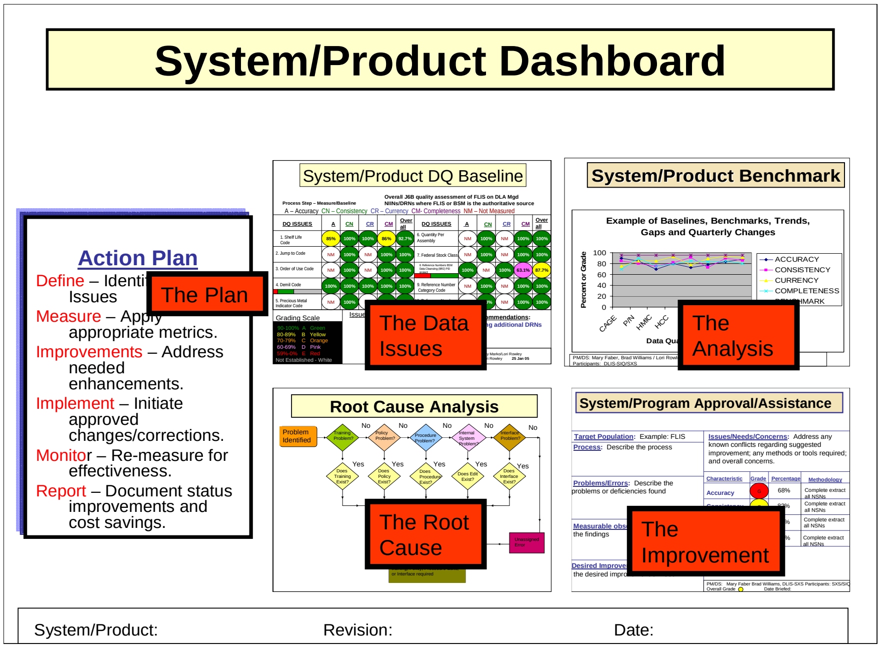
In addition to soliciting consumer feedback, it is necessary to include greater involvement of the data collectors. The PMO discussed the various data quality issues with the focal points that collect the data (Army, Navy, Air Force, and GSA). The original ten issues were selected by the program management office in order to test the methodology. At the quarterly meetings it was apparent that the data collectors had viewpoints that differed from the custodians (program managers and functional managers). The HMIRS FMO realized that consistent attention was needed to continue progress and address additional data quality issues cooperatively; therefore an eight person subcommittee was created. This "small working group" has representation of the various stakeholders in the three categories of data collectors, custodians, and consumers. Representatives from each user group now meet on a quarterly basis to discuss issues and monitor progress. Involvement of all parties has allowed for a more comprehensive study, greater buy-in, and perhaps a more accurate depiction of HMIRS data quality. This systematic process allows for analysis of trends, improvements, changes, or seasonal fluctuations; information that would be useful for marketing, budgeting, and reporting metrics to upper management.

As the data steward, the PMO created a "stoplight chart" during the *measure phase* to illustrate the status of each data quality indicator (refer to Figure 3). These are measured and charted on a semi-annual basis. Although the charts allows for an over-all system grade for data quality, the PMO found this to be somewhat inaccurate and self defeating. The chart allows the reader to drill down to the issue and root causes. Assigning an overall grade would result in a loss of visibility of issues and portray an inaccurate picture of data quality. Upon analyzing the charts and researching root causes, they will provide feedback to the focal points at the quarterly small working group meetings. This would be an opportunity to provide praise, make recommendations for change, and make collective decisions for additions or deletions.

In addition to tracking quality indicators, it is necessary for the quality working group to establish benchmarks for each indicator. The sample graphs illustrate all ten variables having a benchmark of 90% or greater data quality rating. Upon further analysis, however, each variable may have a different benchmark. For example, 80% quality may be adequate for some issues. It is also recommended to assess the return on investment for the issues under review. Although a 90% rating may be desirable, the high cost to improve the quality may not be an effective use of funds. It will be necessary for the stakeholders to assess the cost and feasibility of improving quality when establishing benchmarks. The dashboard chart in Figure 8 illustrates the "big picture" of data quality status. This sample was taken from the DLIS Data Quality Manual and contains generic data; however the template is a useful tool when briefing senior management [25].

Following the analysis, it is also recommended that the results be used to determine whether changes need to be made to the database schema, business rules, and system edits. Although this paper was an introductory study, it reveals that there is a need for greater awareness of the issue of data quality. Next steps should include training for various stakeholders. Educating collectors on correct input procedures may decrease the amount of erroneous data entered into the system. Collectors may not be aware of the implications of poor data quality, and the effect of careless data entry. Training will encourage greater attention to detail and pride in work. An emphasis on quality information will result in enhanced awareness and fewer errors.

Figure 8



System/Product Dashboard

Action Plan

Define – Identi[fy] Issues

Measure – Apply appropriate metrics.

Improvements – Address needed enhancements.

Implement – Initiate approved changes/corrections.

Monitor – Re-measure for effectiveness.

Report – Document status improvements and cost savings.

The Plan

System/Product DQ Baseline

Process Step – Measure/Baseline

Overall J6B quality assessment of FLIS on DLA Mgd NIINs/DRNs where FLIS or BSM is the authoritative source

A – Accuracy CN – Consistency CR – Currency CM- Completeness NM – Not Measured

DQ ISSUES	A	CN	CR	CM	Over all	DQ ISSUES	A	CN	CR	CM	Over all
1. Shelf Life Code	85%	100%	100%	86%	92.7%	6. Quantity Per Assembly	NM	100%	NM	100%	100%
2. Jump to Code	NM	100%	NM	100%	100%	7. Federal Stock Class	NM	100%	NM	100%	100%
3. Order of Use Code	NM	100%	NM	100%	100%	8. Reference Numbers BSM Data Cleansing (BR2) PID project	100%	NM	100%	63.1%	87.7%
4. Demil Code	100%	100%	100%	100%	100%	9. Reference Number Category Code	NM	100%	NM	100%	100%
5. Precious Metal Indicator Code	NM	100%							NM	100%	100%

Grading Scale

90-100%	A	Green
80-89%	B	Yellow
70-79%	C	Orange
60-69%	D	Pink
59%-0%	E	Red

Not Established - White

Issue

[Reco]mmendations:

[...]g additional DRNs

[...]y Marko/Lori Rowley

[Lo]ri Rowley 25 Jan 05

The Data Issues

System/Product Benchmark

PM/DS: Mary Faber, Brad Williams / Lori Rowl[ey]

Participants: DLIS-SIQ/SXS

The Analysis

Root Cause Analysis

The Root Cause

System/Program Approval/Assistance

Target Population: Example: FLIS

Process: Describe the process

Issues/Needs/Concerns: Address any known conflicts regarding suggested improvement; any methods or tools required; and overall concerns.

Problems/Errors: Describe the problems or deficiencies found

Characteristic	Grade	Percentage	Methodology
Accuracy	G	68%	Complete extract all NSNs
Consistency		82%	Complete extract all NSNs
		%	Complete extract all NSNs
		%	Complete extract all NSNs

Measurable obs[ervations]: the findings

Desired Improve[ment]: the desired impro[vement]

PM/DS: Mary Faber Brad Williams, DLIS-SXS Participants: SXS/SIQ

Overall Grade ● Date Briefed:

The Improvement

System/Product: Revision: Date:

Due to the nature and content of the data, there is often pressure for data to be entered as quickly as possible in order for it to be processed and available for the end user, the warfighter. This pressure for data "in the right time" often compromises the other qualities for which DLA strives: the "right place" and the "right item". It is necessary for policies and procedures to address the issues of speed and data availability versus data quality, and then provide awareness training to all stakeholders for consistent implementation.

HMIRS is a diverse database that involves numerous parties. Further research to address various perspectives and concerns is necessary in order to more fully address the issue of data quality. Future studies that entail systematic data extractions, regular measurement and tracking, comparisons of benchmarks and role gaps, and analysis of root causes and affects by the subcommittee, will provide a comprehensive and unified approach to data quality.

LIMITATIONS

It is realized that the methodology is not without fault. The data was extracted from databases that are constantly changing as data is constantly being inputted. Therefore, when extracting data, one may get different results just one hour later even though the same process to extract the data and the same formulas were used. This does not mean that the study is not reliable; however, the researcher must be very careful to fully document the time and date of the data extract and specify the SQL commands used to extract the data as well as how the data was used (i.e., the formula used to make conclusions).

Other vulnerabilities with the methodology included the selection of the ten variables. The data quality issues were selected based on hypothesized quantity of errors, level of importance, level of control, and finally, feasibility of extracting data. The study does not exhaust all data quality concerns, but merely serves as a starting point and a template from which to introduce the topic of data quality to stakeholders and repeat the study using other variables. If other issues were used that entailed different aspects of the system, it is realized that this would produce different results and thus a different aggregate grade. The intention of the study, however, was not to be the end, but a means to the end, a study of existing methods for assessing data quality. The results will be used to initiate further study and justify data quality as being a valid agenda item for stakeholder meetings.

Readers should not assume that the selection of various issues results in a subjective study. As a note of caution, different stakeholders may choose issues that are most important to them when replicating the study. This will produce different results and grades; however, each study may still be valid. This demonstrates the issue that quality has various definitions depending on the user group. It would be advantageous for the HMIRS program management office to perform various quality studies according to the parameters defined by various consumers.

CONCLUSION

Data quality is a growing concern among organizations due to greater information sharing. In the past, IT departments were given the responsibility for data strategies as information management was considered a systems issue as opposed to a product. The most well managed database system, however, will fail to deliver quality data if proper procedures and emphasis on quality do not exist. Risks will continue to increase unless program managers confront the problem and use a multi-dimensional approach that includes input and buy-in from the data collectors, custodians, and consumers. Organizations must also use a comprehensive methodology that includes defining data quality and unique needs, measuring quality, analyzing findings, and identifying and implementing an improvement plan.

This project paper has begun the process using the DLIS Data Quality Plan and the Total Data Quality Management methods as a guide. The TDQM method has proven to be a useful tool with a few adaptations to meet the unique needs of HMIRS. The paper describes the initial steps used thus far and outlines those objectives that are yet to be accomplished. Data quality is an ongoing process. It is the means to an end but not the end itself. Establishing a baseline will assist in strategic planning. Comparing baselines to benchmarks will be used to guide decisions regarding resource allocation and customer relationship management plans. Analysis of trend lines will reveal success stories used for marketing or expose areas of need to justify budget allocation.

It takes courage for an organization to scrutinize its data, as it places it in the spotlight or under the microscope depending on what is revealed. The Defense Logistics Information Service prides itself as the premier agency for defense logistics. As the “information broker” for defense, data is the foundation of DLIS systems. Therefore DLIS, and subsequently HMIRS, must continuously strive to provide quality information for the war fighter in order to maintain that solid foundation.

REFERENCES

[1] Fabiszak, K., & Nguyen, T. (2002) Enhance the quality of your data. Retrieved March 23, 2005 from www.DMReview.com.

[2] Beal, B., (2004) Bad data haunts the enterprise. Retrieved March 23, 2005 from www.SearchCRM.com.

[3] Pierce, E. M. (2003, October) A progress report from MIT information quality conference. *The Data Administration Newsletter (TDAN.com)*. Retrieved June 10, 2004, from http://www.tdan.com/:026hy04.htm

[4] Defense Logistics Information Service (2003). *Year In Review*.

[5] Defense Logistics Information Service (2003). *Year In Review*.

[6] Wang, R., & Madnick, S. (1990) A polygen model for heterogeneous database systems: The source tagging perspective. *16th International Conference on Very Large Databases, August 1990,* 519-538.

[7] Codd, E.F. The Relational Model for Data Management: Version 2. Addioson-Wesley Publishing Co. Reading, MA. 1990.

[8] Wang, R., & Madnick, S. (1990) A polygen model for heterogeneous database systems: The source tagging perspective. *16th International Conference on Very Large Databases, August 1990,* 519-538.

[9] Kovac, R., Lee, Y., Pipino, L. (1997) Total data quality management: The case of IRI. *Proceedings of the 1997 Conference of Information Quality*, *October 1997,* 63-79.

[10] Wang, R. (1998) A product perspective on total data quality management. *Communications of the ACM, 41*(2), 58-63.

[11]Pipino, L., Lee, Y., & Wang, R. (2002) Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

[12] Lee, Y. & Strong, D. (2003-4) Knowing why about data processes and data quality. *Journal of Management Information Systems*, *20*(3), 13-39.

[12] Wang, R., & Strong, D. (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5-33.

[13] Pipino, L., Lee, Y., & Wang, R. (2002) Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

[14] Wand, Y., & Wang, R. (1996) Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.

[15] Lee, Y., Strong, D., Kahn, G., & Wang, R. (2002) AIMQ: A methodology for information quality assessment. *Information & Management*, (40), 133-146.

[16] Lee, Y., Strong, D., Kahn, G., & Wang, R. (2002) AIMQ: A methodology for information quality assessment. *Information & Management*, (40), 133-146.

[17] Lee, Y., Strong, D., Kahn, G., & Wang, R. (2002) AIMQ: A methodology for information quality assessment. *Information & Management*, (40), 133-146.

[18] Pipino, L., Lee, Y., & Wang, R. (2002) Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

[19] Wang, R., Storey, V., & Firth, C. (1995) A framework for analysis of data quality research. *IEE Transactions on Knowledge and Data Engineering*, *7*(4), 623-640.

[20] Funk, J., Lee, Y., & Wang, R. (1998) Institutionalizing information quality practice: The S.C. Johnson wax case. *Proceedings of the 1998 Conference on Information Quality*, October 1998, 1-17.

[21] Defense Logistics Information Service (2004). *Data Quality Manual*.

[22] Defense Logistics Information Service (2004). *Data Quality Manual*.

[23] Pipino, L., Lee, Y., & Wang, R. (2002) Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.

[24] Wang, R., Ziad, M., & Lee, Y. (2001) *Data Quality*. Boston, MA: Kluwar Academic Publishers.

[25] Defense Logistics Information Service (2004). *Data Quality Manual*.