A SYSTEMATIC APPROACH TO ENSURE DATA QUALITY WITHIN EMERGING SYSTEMS

(Research in Progress)

Abstract: This paper describes an update to the Defense Logistics Information Service Data Quality Manual that deals with ensuring data quality for emerging systems. Background information is presented describing the Data Quality Manual with a more complete review of the new content which focuses on the Emerging Systems section.

Key Words: Department of Defense (DOD), Defense Logistics Agency (DLA), Defense Logistics Information Service (DLIS), Data Quality Manual (DQM), Program Management Office (PMO), Data Steward (DS), Commercial off the Shelf (COTS), Government off the Shelf (GOTS), Business System Modernization (BSM), Enterprise Business Systems (EBS), Federal Logistics Information System (FLIS), Inventory Control Point (ICP)

INTRODUCTION

The Defense Logistics Agency (DLA) is a Department of Defense (DOD) agency whose mission is to provide the best valued integrated logistics solutions to America's armed forces and other designated customers, in peace and in war, around the clock, around the world [1]. The Defense Logistics Information Service (DLIS) is a field activity of DLA. Its mission is to provide interoperable, integrated, quality logistics data and enterprise IT solutions for joint warfighters, the Military Services, the Defense Department, other Federal agencies and international partners in order to optimize the effectiveness and efficiency of the DOD supply chain [2]. DLIS manages the catalog system for the Armed Forces. All items of supply, from the smallest nut or bolt to major end items are listed in this catalog. DLIS doesn't deal with the items directly; however it does deal with the data pertaining to these items. With over 100 data elements per item, standardized across all Military Services, it is easy to see why data quality is a must and not just nice to have.

BACKGROUND

In response to DLIS' mission to provide quality logistics data, the Data Integrity office was established as a means to create an organizational focus that is centered on data quality. The Data Integrity office utilizes two primary methods to identify data quality issues. First, are data issues identified by various sources (typically customers of our data) and second, system reviews as governed by the DLIS Data Quality Manual (DQM). The DQM is the rule and guide for all system reviews to check for possible data quality issues. In addition, it provides guidance on the roles and responsibilities of both the Data Integrity office Data Stewards (DS) and the Program Management Offices (PMO) located in various areas of DLIS. By going through a six step process, measuring by four key characteristics, and using the various roles and responsibilities created for each step, data quality is reinforced for the systems. The DQM has been recently updated with a section that addresses data quality issues in new systems that are under development. This new section explains how to foster data quality in emerging systems and follows one underlying principal: it is better to build in data quality up-front than to try to bolt it on after the fact.

THE DATA QUALITY PLAN

Before looking into data quality for emerging systems, we need to look at the basic data quality process DLIS uses for current systems as shown in the Data Quality Manual. The DQM is made up of multiple sections, but this paper will only address the areas which deal with the current system process and then the new emerging systems section. This will provide the readers with a comparison of the approaches used to identify and correct data issues between a current data system and an emerging system.

The Process

This section of the manual details the six step process involved in a DLIS data quality review of an existing system. This is visualized by the graphic below (figure 1). A circle was used to show that this is a repeated process that never truly ends as long as new issues are reported and the system is still in use. It is the DS and PMO job to fulfill the roles and responsibilities assigned to them at each step in the process. Each step of this process is described in more detail along with defining the roles and responsibilities of the DS and PMO.

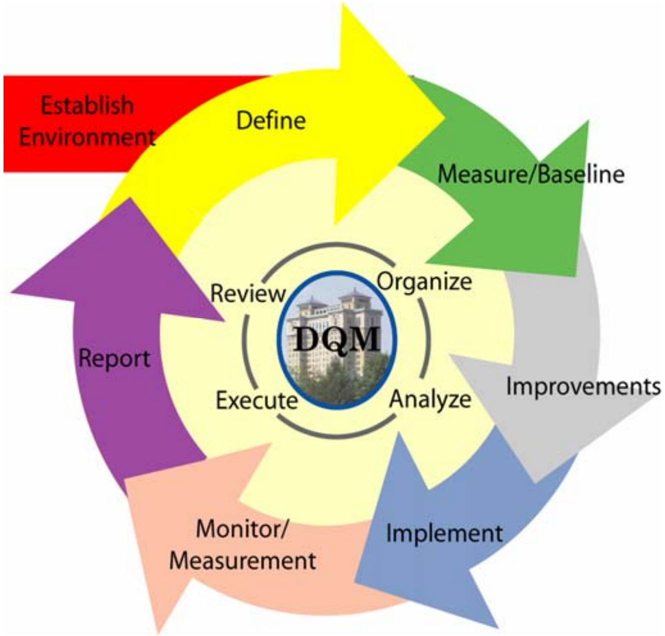


Figure 1 – Data Quality Process

Step 1: Define

This step starts the system review process by focusing on the system itself as well as any data quality issues that are already known to be inherent in the system. The DS starts out, with the help of the PMO, by creating an action plan for the system review. This plan is a document that specifies the players involved, defines the current system environment and ultimately provides a roadmap for all six steps of the process. By its nature, this plan is only partially completed at the start and more information is added as each step is fleshed out and individualized for the system that the plan represents.

Some of the other tasks associated with the define stage are:

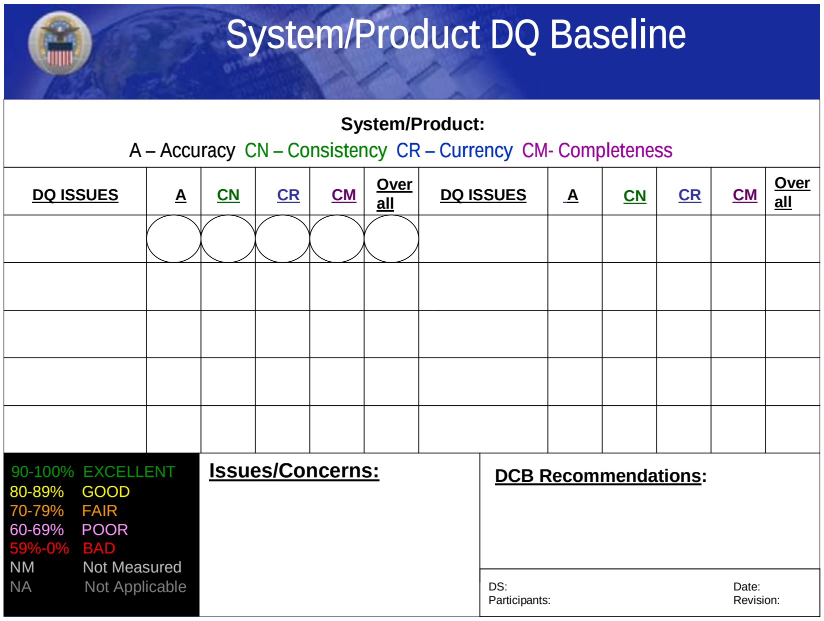
- determine the scope of the review and define the actually data quality issues
- review of all process guides, instructions and data flows for the system
- obtain extracts and raw data needed for review
- development of a system product roadmap

It is important to note there is no set limit to the number of issues that can be covered during a system review and no set timeframe to begin reviewing them. As new issues are uncovered they can be added. What this means is that even though the overall system review may be at a later step in the six step process, new issues can enter through the define step at any time. The key factor here is to have them clearly defined.

Step 2: Measure/Baseline
Once the define step has been completed, the data quality issues have been identified, and data is available for measuring, the DS and PMO will begin analyzing the quality of the data in regards to four DQ characteristics described below:

- **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 system edits regarding length, format, and accepted values (including combination edits).
- **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.

Based on these characteristics, the DS will select the size and type of data to be measured, apply the appropriate metrics, and analyze the results. From this data the results can be shown a couple different ways. The most common is through the use of a system product baseline chart that uses a percent to show how each issue measured up based on the metric(s) used (see figure 2 and 3).



System/Product DQ Baseline

System/Product:

A – Accuracy CN – Consistency CR – Currency CM- Completeness

DQ ISSUES	A	CN	CR	CM	Over all	DQ ISSUES	A	CN	CR	CM	Over all
	○	○	○	○	○						

90-100%	EXCELLENT
80-89%	GOOD
70-79%	FAIR
60-69%	POOR
59%-0%	BAD
NM	Not Measured
NA	Not Applicable

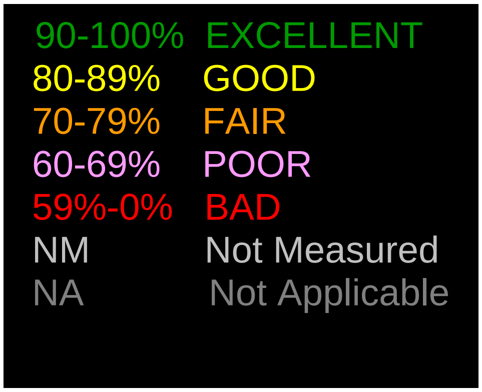
Issues/Concerns:

DCB Recommendations:

DS:
Participants:

Date:
Revision:

Figure 2 – System Product Baseline Chart



90-100%	EXCELLENT
80-89%	GOOD
70-79%	FAIR
60-69%	POOR
59%-0%	BAD
NM	Not Measured
NA	Not Applicable

Figure 3 – System Product Baseline Chart Percentage Block

The chart is used to identify the DQ issues and each metric circle is filled in with the color corresponding to the percent score for that issue against its metrics. Inside the colored circles, then, the percent is given as a number to allow the reader to have both a numeric and a color queue. Some issues will use all four metrics and others will only use a couple as appropriate to those issues. The determination for which metrics are appropriate will be made during the define step between the DS and the PMO. This information is then provided as a tool for management and the PMO to gauge what needs to be done and as a means to determine top data quality priorities for resolution.

Step 3: Improvement
After the data is collected and a comprehensive root cause analysis is done on the findings in step 2, step 3 will begin with a review of the findings by the PMO. It is the PMO's responsibility to determine what improvements or corrective actions to pursue. There is a wide variety of approaches during this step. Many different techniques can be used depending on the type of problem, the severity, and PMO preference just to name a few. The DQM provides a System/Product Improvement/Assistance Chart to assist the PMO in making some of these determinations.

Step 4: Implementation
During this step, the PMO will take the necessary action to obtain tools required for all approved improvements from the last step. Some issue fixes will take more time or cost more to implement than others. A programming change to fix a technical issue may be required, whereas a constant input error may require training of data input to personnel. The PMO will ensure implementation completion by establishing timelines and milestones and will document everything within the Action Plan.

Step 5: Monitor/Measure
This particular step has its focus shifted back to the DS as the primary person responsible for actions. It is the job of the DS during this step to continually monitor the implemented improvements to the data. Considerable time may elapse before improvements are realized. A mass change within the system could show immediate improvement, however training personnel and updating procedures for data entry may take longer slowing improvements to data quality.

Step 6: Report
This step primarily deals with reporting of progress of data quality improvements. Based on policies and procedures of a particular organization, various reports may be required to provide the information necessary to measure data quality improvements. It is therefore the purpose of this step to ensure the proper reporting policies are followed with the reports provided to the appropriate audience.

THE NEED FOR A DATA QUALITY PLAN FOR EMERGING SYSTEMS

Over the past few years DLA has instituted its new Business System Modernization (BSM) system, now referred to as Enterprise Business System (EBS). During BSM concept demo some data quality issues were identified early on. A data cleansing team was created to look into the various data issues and identify areas where attention was most needed. For example, the team first identified records that were candidates for cancellation and then coordinated with the customers for their concurrence. Part numbers was another area where data issues were identified. Over 2.6 million part numbers were reviewed with 33% requiring an update.

DLIS was actively involved early in the Business System Modernization data cleansing effort because the DLIS managed Federal Logistics Information System (FLIS) shares 48 common data elements with BSM. In June of 2003, DLIS briefed HQ DLA on the need for developing a plan for BSM data cleansing. Working with the HQ DLA lead, DLIS was assigned the role of trusted agent. Together with the DLA Inventory control points, a total of 14 data cleaning business rules were developed. Based on these established business rules, DLIS was given authority to update specific data elements in FLIS. DLIS Catalogers reviewed over 1.8 million records with approximately one third resulting in data correction. DLIS also tracked progress on all 14 Business rules to ensure the action was completed.

Overall, BSM was a learning process for DLIS. It was the biggest emerging system to roll out in decades and we were a major part of it. Based on our success in assisting with the BSM launch, DLIS has now taken an active role in other emerging systems. Even with the resounding success of BSM, no process exists in a vacuum. DLIS learned a great deal from its BSM experience and it was determined that the basic process concepts should be captured in the DLIS DQM. Some amount of flexibility is required in the process when reviewing data quality since every emerging system is different. What follows is the Data Quality Manual process for Emerging Systems.

THE DATA QUALITY PLAN FOR EMERGING SYSTEMS

The data quality plan for emerging systems is very similar to the plan for existing systems but it does have some differences. To make it easier for the reader to go from the section on the Data Quality Process to the Emerging Systems section, it is referenced in stages similar to the steps of the previous section. These stages are preparation, data analysis and business rules, data cleansing, progress tracking and sustainment. These stages are shown as a linear progression instead of a circle. Once an emerging system goes live, any data quality initiatives then fall under the basic data quality process referenced earlier in this paper. This means that unlike the earlier data quality plan referenced this does actually have an end point.

Emerging Systems

An emerging system is any new large scale initiative utilizing modern technology (i.e. Commercial off the shelf (COTS) or Government off the shelf (GOTS)) to replace a legacy system. The data quality review for an emerging system is somewhat different from that of an existing system and is better labeled as a data cleansing effort. In this review, all the data residing in the legacy system and any interfacing systems will be cleansed for Accuracy, Currency, Consistency and Completeness prior to conversion to the emerging system.

The data cleansing process for emerging systems consists of five stages: Preparation Stage, Data Analysis and Business Rules, Data Cleansing, Progress Tracking, and Sustainment. This process is very similar to the six step process performed on System/Product DQ Reviews, previously outlined in an earlier section of the Data Quality Manual. A comparison of the two processes is provided below in Figure 3 Data Cleansing Process Compare.

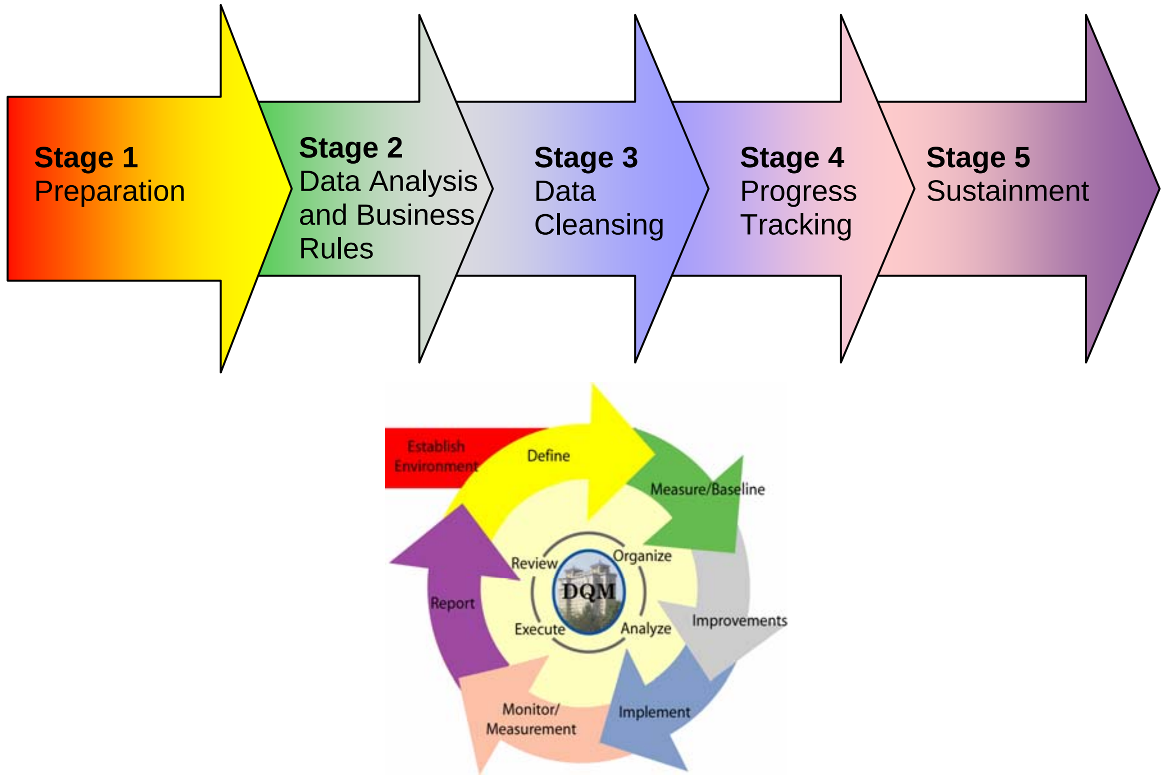


Figure 3 - Data Cleansing Process Compare

Stage 1: Preparation

A data cleansing team is established consisting of subject matter experts from the interfacing system, the emerging system PMO and the Data Integrity Office. A data cleansing team lead will be appointed to conduct meetings and ensure all members of the team are kept abreast of the data cleansing process to be performed.

Once the data cleansing team is established, the team will review the requirements of the emerging system and begin identifying data elements that will reside in the new system. A roadmap will be created to show these data elements and where they currently reside.

After the roadmap is developed, the data cleansing team will document priorities and expected performance standards of the emerging system. The following are examples of concerns that may need to be addressed:

- What primary data elements will be in the first release?
- What will good data look like as opposed to bad?
- Are the data elements in the interfacing systems going to show up exactly the same in the emerging system or will field names, acceptable characters, and definitions be changing?
- What data quality issues exist in the interfacing systems that may pose a problem to the emerging system?

Stage 2: Data Analysis and Business Rule Creation

The data cleansing team will identify data issues (data problems) and compare the data against the four data quality characteristics, Accuracy, Consistency, Currency and Completeness.

The following are additional concerns that may need to be addressed when reviewing the four data quality characteristics:

- First, the team will review the data elements for accuracy, to see if the data is being represented correctly.
- Second, the team will review the data for consistency. In theory, the emerging system will perform or process data differently than its legacy system. Field lengths, what the data elements actually mean and how they are presented may cause differences during the conversion process. A review of consistency is therefore two fold. The data element must be consistent among all the interfacing systems and must convert into the proper format in the new system.
- Third, the team will review the data for currency. When multiple systems are involved, special care will be taken when determining which system contains the current data. Reviewing outdated data is not practical unless the emerging system itself is simply a historical data warehouse.
- Finally the team will review the data for Completeness. The team will ensure all of the data required from the legacy system makes the transfer to the emerging system. If the emerging system requires a certain data element to be mandatory but the legacy system did not, then there will be incomplete data that will need to be generated by other means.

Based on the data issues identified during this stage, the team will develop a root cause analysis to determine whether the causes are systemic or manual. This will help determine what corrective actions are needed. Because systemic corrections can be done easier and quicker, the PMO may decide to take care of these problems first.

Once the team has identified the issues and causes, business rules will be developed to provide the standard for the data elements that the new system is trying to obtain. Business rules show what is a good value for that data element, who will fix it and how it will be fixed.

The team will create business rules for accuracy and for the conversion (consistency and completeness) of data. This may require the team to develop new system edits, as well as mock conversions to ensure the data will transfer correctly.

A mock conversion is run in a test database. It utilizes a snap shot of the data provided by an interfacing system. The data is run through an identical process as if it were going into the production database. Afterwards, the data is checked and any discrepancies identified are reviewed. These mock conversions can identify new problems and help refine the business rules as needed.

Stage 3: Data Cleansing

In this stage the team will take the necessary actions to implement improvements/corrective actions. This stage is very similar to the implementation step for a System/Product DQ review, however, the corrections being implemented here follow business rules and are usually accomplished as releases. Because the data corrections will be accomplished in releases, implementing improvements will be repeated as needed and adjusted to the scheduling of each release. As each new release is prepared, new improvements may need to be made against the existing business rules or new rules may need to be created.

A number of errors uncovered during the review of the interfacing systems will be errors in the pre-existing data. These errors need to be corrected by the authoritative source prior to the data being released to the emerging system. This is especially important if the interfacing system will be a continuous feeder of data to the emerging system. Not all possible errors in a given release can be determined in advance. The team will need to analyze the complete release once it's in the emerging system. The team will utilize automated queries and reject notification edits to obtain the identified rejects.

The team will determine if any new business rules need to be added or if any of the ones previously drafted need to be amended to address unique or unexpected situations. New or amended rules must be implemented to the data quality testing and cleansing practices. The team will also need to determine if any special cleansing efforts are needed for data that has already made it into the emerging system.

Stage 4: Progress Tracking

The team will determine what types of reports are required to track the results of each release and provide status to project sponsors and management. The team may decide to utilize a generalized reporting structure for reporting purposes. There are actual examples of this structure in the DQM. This will require a number of charts or graphs requested by the PMO to show the general progress of the data cleansing effort. Reports should include status such as:

- Overall actions.
- Breakdown of actions per systems.
- Balance left to be cleansed.
- Items added or deleted.
- Items that failed to meet a business rule and therefore did not make it into the new system intact.
- Actions that need to be added or deleted to another release.
- Actions that need to be set aside that set of information for further review and instruction.

The team should be prepared for unexpected situations that may occur during a release and be able to adjust their reporting technique to reflect the applicable information.

Other reporting documents such as the System/Product DQ Baseline Chart or the System/Product Action Plan may also be utilized for reporting purposes.

Establishing timeframes for reporting is just as important as the reporting criteria and may be altered throughout the life cycle of the releases. At the beginning of the project the team will establish timeframes for the reports to coincide with the major accomplishments of system development. For each release, the timeframes for reporting may be re-adjusted accordingly.

Stage 5: Sustainment

Once the final release is complete, there may still be work for the data cleansing group to perform. Due to the press of time or resources some errors may still exist from earlier releases. These errors may not have been considered substantial enough to fix at the time, but need to be revisited now for possible corrective actions.

Based on the established business rules as well as the internal edits of the new system, the team will also perform testing on the data to ensure that the standards were met. A series of follow-up testing should be performed and documented to ensure everything is as it should be.

Finally, once the system is in sustainment, it can be added to the list of Targeted Systems/Programs scheduled to undergo a DQ System/Product review [3].

CONCLUSION

Data quality has grown to become the forefront in thinking with organizations responsible for data and data systems. In the past, data quality was considered more as an after thought and it has not been until recent years that the importance of data quality has been realized. In its infancy, data quality initiatives focused on identifying and correcting erroneous data in existing data systems. Processes have been put in place to constantly check data and correct data through system enhancements and/or training with procedural changes.

The next challenge for data quality is with emerging systems that are under development. Typical emerging systems are developed to replace outdate existing systems with the intent to migrate data to the new system. In a worst case scenario the data from a system with poor data quality is migrated to the new system perpetuating poor data.

Originally, the DLIS Data Quality Manual addressed the after-thought process of dealing with bad data resident in pre-established systems. More recently, various initiatives within DLIS and DLA have provided the opportunity to address poor data quality before it infects new emerging systems. The DLIS Data Integrity office, in conjunction with many other parts of DLIS and DLA, performed a comprehensive data cleansing initiative that developed many of the process stages used in the Emerging Systems section of the DQM. In effect, through the BSM/EBS initiative and current initiatives to develop new systems the process has been captured and documented. Obviously no two system initiatives will be the same, and some variations will be necessary, but the general idea of data quality for emerging systems within DLA is now a referenced reality.

REFERENCES

[1] Defense Logistics Agency. (2007) Retrieved from the World Wide Web on 14 June 2007 from www.dla.mil

[2] Defense Logistics Information Services. (2007). Retrieved from the World Wide Web on 14 June 2007 from www.dlis.mil

[3] Defense Logistics Information Service, (2006). *DLIS Data Quality Manual*, Battle Creek, MI, (2-10).