You Need Clean Data

Higher data quality standards are needed to improve Marine Corps decision making

by Maj Amber Coleman

The Marine Corps collects and maintains a significant amount of maintenance data, including the daily usage at using units as well as annual depot maintenance repair costs. However, data quality issues result in analysts spending almost 80 percent of their time cleaning and preparing data sets for analysis instead of transforming that data into actionable intelligence.[1] The time-consuming requirement of cleaning data is costly and is caused by the poor quality control of data going into Marine Corps data collection systems. Decision makers, analysts, and managers at all levels must adapt to accommodate this extra time in their everyday work.[2] If one month is required to develop a working model, an analyst could spend an average of four months cleaning and preparing that data, and there is no guarantee that the analyst will remove all of the erroneous entries.[3] This is the equivalent to spending 80 percent of your time arranging your rifle cleaning gear and only 20 percent actually disassembling, cleaning, and reassembling your weapon. While arranging your cleaning gear is necessary, it should only be a small part of the process compared to the time spent scrubbing and cleaning your weapon to ensure it functions properly. Poor data quality is an analyst's worst enemy[4] as it continues to prevent the Marine Corps from gleaning actionable information from our maintenance data.

In 2014, the Marine Corps Operations Analysis Directorate attempted to study the feasibility of creating a maintenance data collection MOS similar to the aircraft maintenance administration specialist MOS 6046. Ultimately, the MOS was not created, and the Marine Corps chose to simply document the effort to collect maintenance data. The M1A1 and the Medium Tactical Vehicle Replacement (MTVR) were used as test cases. Timelines to collect data varied; while some data was available within days, other data sets never materialized. Of the data collected, approximately one-third was unusable because one of three key fields—serial number, date opened, and defect code—was missing from maintenance documentation. Additionally, there were 397 M1A1s reflected on Marine Corps supply records at the time; however, 1,224 serial numbers appeared in this data set.[5] There were approximately 2,900 MTVRs on the supply records, yet over 6,800 appeared in the data sets provided.[6] Accurate serial number reporting and accountability is the minimum requirement in this effort. Without it, there is no way to tie maintenance actions to specific assets and, therefore, no way to uncover information from that data to identify usage patterns that may lead to predictive maintenance capabilities. It is as if both the maintenance action and the effort to document that maintenance action never happened. Imagine spending hours cleaning your weapon only to find that the armory did not maintain accurate serial number accountability so there was no record of your efforts. Even worse, there might be no record of your weapon being in the armory at all.

The M1A1 was used to collect maintenance data. *(Photo by Cpl Kevin Payne.)*

Marine Corps Logistics Command conducted a study to calculate the maintenance costs to the Operating Forces for each year the Marine Corps deferred AAV depot-level maintenance.[7] The analysts found that six years is the optimal depot maintenance interval, which analytically validated the current AAV depot maintenance strategy. However, attempts to apply the same analysis to other vehicle types were unsuccessful primarily because vehicle serial numbers did not match across various data systems. In 2016, Program Analysis and Evaluation, Programs and Resources,

>Maj Coleman is assigned to Combat Logistics Battalion 6.

HQMC conducted a study to determine the divestment criteria for HMMWVs.[8] Their study revealed patterns between usage data and maintenance histories, but this was based on only 58 percent of the available data. Because of mismatched serial numbers between Global Combat Support System Marine Corps and Transportation Capacity Planning Tool, 42 percent of the data was unusable. This is a critical issue because this missing data might hold key information and contain trends that are absent in the usable data. We can only expose these trends through the data itself, and as of now, there is not enough information to provide accurate predictions.

In 2015, a Naval Postgraduate School student, Maj Adam Foley, attempted to analyze MTVR maintenance trends, but instead found that over 50 percent of the available data was unusable because of missing mileage.[9] Mileage, hours, and any other type of Equipment Operating Time Code (EOTC) data provide a means to determine the age of an item. Without usage data, it is impossible to accurately determine how aged the item really is; thus, there is no way to associate maintenance occurrences with usage trends.

Fortunately, the Marine Corps is not alone. Industries worldwide are attempting to gain further insight from their data, and many suffer from the same problems. One study suggests that only three percent of businesses have acceptable data quality levels.[10] IBM estimates that poor data quality cost businesses over $3.1 trillion in 2016 alone.[11] The best way to improve data is to prevent errors from ever entering the system to begin with.

Data Quality Is Every Marine's Job

Data quality begins at the point of entry—the Marine on the shop floor. These Marines must understand that keeping this data accurate and clean is equally as important as keeping your weapon clean. It consumes no extra resources other than the few seconds it takes to ensure we capture information accurately. This effort will enable the Marine Corps to provide quantifiable and defendable data to support requirements at all levels. Regardless of the systems the Marine Corps chooses to record and archive this data, every Marine has a responsibility to input quality information and work with the tools we have.

Marine Corps analysts currently leverage machine-learning techniques using automated processes to sort through large data sets to find patterns and connect that data with predictable outcomes.[12] Essentially, the machine learns the behavior of your process to provide useful insights and predictions. Based on historical data, analysts may also build mathematical models to calculate risk; regardless, the data is the foundation of this capability. For example, a squad preparing for a patrol could select vehicles and weapons based on the probability of breakdown for each item to increase the overall probability of mission success. Incorporating a feedback loop at the conclusion of each mission provides additional data and enhances this capability since analysts may iteratively improve their models over time as more data and outcomes are collected.

Many of these models, once developed, can run on government networks using open-source software, and the Marine Corps already employs active duty and civilian analysts capable of developing these models at no additional cost to the government. Reducing the confounding "hidden data factory"[13] that constantly operates to link and clean disparate, dirty data will result in more of these analytical resources being available to focus on machine learning and predictive analytics leading to actionable insights. This work will ultimately enhance our understanding of the capabilities and limitations of our equipment before they are needed in combat.

Equipment operating codes are critical to maintaining equipment at a high level of readiness.
(Photo by LCpl Isabella Ortega.)

You Can Do Your Part

The data creation and upkeep is not the sole responsibility of the Marines on the shop floor or the data analysts. Leadership at all levels has the responsibility to maintain data quality through regular data audits. Begin with focusing on just a few data fields such as serial numbers, EOTC data, defect codes, and dates opened and closed. These fields are the most vital to maintenance data and without them data entries are useless. To maintain an understanding of your unit's data quality score, conduct regular in-house data assessments which is much easier than you think.

Managers at all levels could implement the Friday afternoon measurement method.[14] Pull your last 100 maintenance and supply transactions, gather two or three subject matter experts on a Friday afternoon to review each transaction and mark obvious errors.[15] For example, highlight serial numbers from

We have to pay attention to what is happening now, so that we will be prepared for the future.
(Photo by LCpl Isabella Ortega.)

maintenance transactions that do not match your supply records, empty or illogical EOTCs (look for mileage entries such as 12,345 or 99,999), missing defect codes, and empty or illogical dates. Then count the number of errors in each category of data and subtract that from 100. This provides a data quality score for each data element. If scores are high using these variables, begin including more data fields to further increase data fidelity. This methodology is simple and tailorable to any size or type unit within the Marine Corps, making it a low-cost tool that you may periodically employ, ensuring your unit is paying attention to data quality. Conducting this process during Friday afternoons prevents interference with other battle rhythm events throughout the week.

High operational tempo compels us to pay attention to what is happening in the present rather than thinking about how our actions (or inaction) will impact operations in the future. As a result, commanders and leaders at all levels must espouse the importance of data quality just as they underscore the importance of clean weapons. Clean data may not immediately keep you out of danger, but when appropriately leveraged, it could keep you from breaking down in harm's way and potentially save the Marine Corps millions of dollars.

The Marine Corps cannot continue to grow and innovate without keeping better data and ensuring that data works for the institution in a low cost and efficient manner. This effort does not necessarily require more funding. It simply requires education, diligence, and organizational discipline ranging from the shop floor to all levels of leadership. Everyone needs to understand the relationship between the data they are recording and the capability that accurate data may one day provide.

Several civilian and government agencies already capitalize on detailed analysis of maintenance and cost data. They are able to accurately break down costs, requirements, or other data points to provide detailed predictions that justify future requirements and may eventually result in greater profits. Advanced information technology systems could help, but only after we implement the proper education and processes to support accurate data collection.

If we want to be an innovative and advanced fighting force, we must embrace big data and start enforcing data quality standards throughout the Marine Corps. Our data must be as clean as our weapons.

Notes

1. Thomas C. Redman, "Bad Data Costs the U.S. $3.1 Trillion Per Year," *Harvard Business Review*, (Watertown, MA: Harvard Business Publishing, 2016).

2. Ibid.

3. Thomas C. Redman, "If Your Data Is Bad, Your Machine Learning Tools Are Useless, *Harvard Business Review*, (Watertown, MA: Harvard Business Publishing, 2018).

4. Ibid.

5. Operations Analysis Division, Marine Corps Combat Development Command, "Cost of Data Efficiency," (Quantico, VA: 2013).

6. "Cost of Data Efficiency."

7. Brian Bagley, et al., "Cost Drivers in Vehicle Maintenance An Analytical Perspective," *Phalanx*, (Arlington, VA: Military Operations Research Society, 2016).

8. Randal Cole, Andrew Mathis, and James Bagg, "Resources to Readiness: Equipment and Fiscal Data Analysis," (presentation, Operations Research Operational Advisory Group, Quantico, VA: December 2016).

9. Adam Foley, "Data Quality and Reliability Analysis of U.S. Marine Corps Ground Vehicle Maintenance Records," (masters thesis, Naval Postgraduate School, 2015).

10. Tadhg Nagle, Thomas C. Redman, and David Sammon, "Only 3% of Companies' Data Meets Basic Quality," *Harvard Business Review*, (Watertown, MA: Harvard Business Publishing, 2017).

11. IBM, "The Four V's of Big Data," IBM Big Data & Analytics Hub (Online), available at http://www.ibmbigdatahub.com.

12. John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy, *Fundamentals of Machine Learning for Predictive Analytics*, (Cambridge, MA: The MIT Press, 2015).

13. "Bad Data Costs the U.S. $3.1 Trillion Per Year."

14. Thomas C. Redman, "Assess Whether You Have a Data Quality Problem," *Harvard Business Review*, (Watertown, MA: Harvard Business Publishing, 2016).

15 "Assess Whether You Have a Data Quality Problem."