The Hidden Enemy - Dirty Data

Stephen J. Thomas | Subject Matter Expert

In the world of maintenance and reliability, problems confront us daily. Some of these become part of the computerized maintenance management system in the form of a repair work order. Others termed "emergency work" get the immediate attention of the workforce to quickly correct the problem. But out there in our world lurks a hidden enemy; dirty data. We confront it every day, but when the enormity of the problem is recognized we are unable to address it because the effort to fix it is overwhelming. Consequently we develop group and individual workarounds to mitigate the problem and over time we tend to forget that the problem even exists. The term for this is developing a "scotoma"; the partial loss of vision or a blind spot on an otherwise normal field of vision. In other words we know the problem is there but individually and collectively we have figured out ways to avoid addressing it. For most decisions our workarounds suffice especially in cases that do not involve other individuals or organizations. That is not always the case. Where decisions span multiple departments or job functions working with dirty or inaccurate data is a serious risk to safety, environmental, health and in general operation of the facility.

There are many reasons that virtually every facility has data integrity problems. Here are just a few;

- Multiple databases exist with duplicate information about our assets and no clear recognition of which set of data is correct
- Some databases have too much information which over time can't be maintained and essentially is unreliable
- Some databases have too little information so our ability to make reliability based decisions is virtually impossible
- Often incorrect data is entered into various databases causing serious problems down the road
- There is no feedback loop from work completed in the field to keep the data in the system correct or update it as changes are made to the assets

- There is lack of accessibility to the data by those who needed to make decisions causing them to find alternatives
- Individuals create personal data files as they discover that the information in the systems is often in error. This is "uncontrolled documentation" and has the potential to cause serious problems
- A disconnect exists between the completion of engineering projects and the information provided for entry into the maintenance management system. This should be covered by the MOC process but often is overlooked in an effort to complete and close projects on schedule
- Some systems allow far too many people to be able to alter the data resulting in a total lack of data governance and controlled work process.

I believe if you take a careful look at the asset related data in your system that you will find at least one or more of the above bulleted items describing your situation. However, recognizing the problem is only the beginning. The question is what can you do to correct this serious issue within your data? It isn't going to be easy and requires the effort of every organization involved with your plant assets. So here is a high level approach of to how to go about fixing the dirty data situation.

Assessment

Plants that I've worked in have asset counts in the tens of thousands. What you don't know is where all of the asset related data is stored, and for each the quality of the information contained therein. So your first step is to identify the size of the problem. It's virtually impossible to assess the quality of the data for thousands of assets in multiple databases. There is light at the end of the tunnel because you can gain the same level of information by taking a large statistical sample from the various databases to determine the level of quality. There are consulting firms that can accomplish this task for your organization. They have systems and expertise to do this in the most economical fashion and provide you with an accurate snapshot of the level of your data integrity. I would not suggest trying to do this yourself simply because everyone has a day job and probably very limited time to focus on an effort of this sort. The consultant will give you their full attention and return results that accurately reflect the overall data integrity problem.

Cleanup

There are two options to cleaning up dirty data. Let's call the first, the "Big Bang" approach. This approach involves hiring a consultant firm to actually walk down all of your plant equipment and validate the data within your plant databases. This is a very expensive process but it gets the work done in a timely fashion. I personally was involved with this approach for several refineries for which I worked and in each case the cost exceeded one million dollars. While the effort gets completed in a timely fashion, the cost is excessive.

The second approach is to fix the asset related discrepancies as you go. This requires a very closely monitored work process in order to assure that every time the technicians or operators are involved with an asset that they validate the data within the database. Additionally someone needs to be tasked with making the corrections in a controlled fashion. This approach will take years but is extremely low-cost. The problem is with making this a sustainable process.

There is one other issue that you need to address as part of the cleanup process. You can't have multiple databases with the same data if the data is not synchronized. Without the synchronization aspect, it is far too easy for data to be entered for the same attribute into the different systems differently. The problem related to this is that you never know which of the data elements is correct and a wrong decision could be costly. There are ways to fix this problem. First is to build a system interface that ties all of the different databases together with one database being primary. This is expensive and requires constant updating as your applications change. Since you most likely have numerous and similar databases a Digital Data Management System (DDMS ™) can better address this problem without having to reconfigure your existing applications. The DDMS can also be set up to flag attributes that are not aligned. The other less attractive effort is to have a single database administrator who is responsible to monitor the databases and assure that the information is always synchronized.

Sustainability

Assuming that you embrace the process of cleaning up and synchronizing your asset related data, the next and most important step is to assure that all the work that you have accomplished is sustained over time. While your organization can write procedures, protocols and develop work processes, efforts such as sustaining data integrity over the long-term are very difficult and often as the years pass fall into disrepair.

The solution to data integrity sustainability is data governance. While most do not want to hear it, this is a full-time job. For engineering projects it requires closely monitoring the MOC closeout process to assure that the correct data is getting into the system. For all other changes affected by Maintenance or Operations (also covered by the MOC process) there needs to be in place rigid procedures. In the end its fine to identify your problem and work to clean up the data integrity issues, but without strict tightly controlled governance process your efforts over time will be in vain.

Conclusion

Cleaning up a plant's data integrity issues is not simple and not inexpensive, but failure to do so leaves in place various workarounds for erroneous data driven reliability based decisions. The problem is that these workarounds don't always lead to a successful outcome and in many cases could have a serious impact on your facility.

I'm not suggesting, nor providing the details required to fix this problem. What I am trying to do is to get you to recognize that this problem exists and get you to think about ways to correct it. This will go a long way to assuring that your decisions yield safe, environmentally sound and reliability based actions.