



Data Quality Happyland: How To Get There?

Tom Breur
March 2009

Introduction

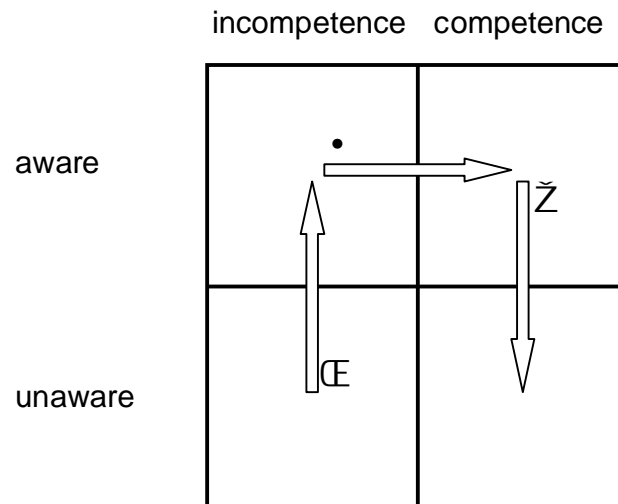
Many organizations are suffering from poor data quality. Some know about their issues, but for many the extent of costs they are incurring due to data non-quality remains largely hidden¹. Since nobody is happy with poor quality, the ubiquitous question is: "How do I accomplish *sustainable* improvement in data quality?"

The reason this is such a challenging question lies in the fact that every company needs different measures, a unique approach. One size does not fit all. But although there are so many differences, there is nonetheless a structure, a pattern that can be discerned which predictably leads to improvement. In this paper we'll expose a simple and familiar framework to guide your progress. Identifying where you are in this structure helps you determine where your highest leverage points for change are. And also, how you can make your improvement effort *last*.

A familiar framework : awareness and competence

For this paper we will use a familiar framework to describe stages that companies go through when they embark on data quality improvement efforts. It consists of a classic 2 × 2 matrix crossing awareness and competence.

Figure 1: organizing for data quality



There is a natural progression moving from the bottom left quadrant via the top left corner to the top right, and then on to the bottom right. Along the way different efforts are required (arrows ϵ , \bullet , and \tilde{Z}).

Inform

Many companies have data quality issues that they are largely unaware of. At least, *most people* within the organization are unfamiliar with their data quality issues, and the (downstream) costs they are incurring as a result of this. The way to get "out" of this place, is by information (ϵ). Tell as many people as possible about the issues you are faced with, and preferably also what the costs are that are associated with the current status quo².

Certainly senior management should be informed about your data quality problems. Translating the consequences of data non-quality into the one dimension that every manager in every industry understands so well, namely dollars, is a great way to draw their attention. Such an effort will require hard work and making assumptions to finalize the cost calculations. Then include the calculation model and assumptions being made. Maybe you want to present this with a certain bandwidth to acknowledge the uncertainty in your calculations. But make sure "a" number gets calculated. Financials have a tendency to "stick", to be remembered quite well.

Educate

The second transition you will want to go through, is moving from awareness/lack of competence to awareness/competence (•). The way you get there is through education. With this transition your goal is to train people in practices that will prevent poor data from entering your systems.

At the front-end, this could be by training data-entry staff in best practices, like always having “four eyes” check manual entry. Maybe you want to reconfirm the importance and cost associated with poor quality (that was raised to their attention in the previous phase), in particular as a standard practice for new trainees. But it could also mean simply designing better user interfaces that enforce and facilitate higher quality data-entry.

On the back-end it means showing how data warehouse staff can track and establish data quality, preventing poor quality data from entering your data warehouse. This might be, for instance, putting deduplication technology in place that leads to fewer duplicate records from the ETL process. Or, reporting the number of errors occurring in your audit dimensions³.

Rethinking accountabilities

The third stage in this voyage is meant to solidify new working practices. This phase should ensure that the new practices become ingrained in the organization and producing high quality data becomes the norm⁴. You accomplish this by restructuring accountabilities (Ž).

Issues you will be facing here are things like organizational alignment and performance targets. If you were rewarding data entry staff for speed, you will now need to add performance objectives that *also* reward staff for making fewer errors. Otherwise, you put them in an unfair quandary when they are overloaded with work, and the only way to deliver quality would be by missing their productivity targets.

Misalignment *between* departments occurs when the people suffering from lack of data quality can't influence resource allocation where (upstream) data quality should be produced. For instance, marketing often incurs the costs as a result of sloppy data handling by data-entry staff. The problem holder is person “suffering” from poor data quality,

in this case marketing. The problem owner controls the resources needed to resolve the problem, in this case the manager of data entry. Organizational alignment is the result of bringing problem holder and problem owner as close together as possible.

The journey continues

After you have gone full circle, new and improved levels of data quality have become the norm. You will have controls in place, and awareness about the importance of quality operations will continue to grow. What then happens, invariably, is that new and heretofore ignored areas will become object of scrutiny. For any other process where the possibility exists to drive out non-quality you will run through this loop again.

As the process of improving data-quality becomes increasingly familiar, you may attempt to do several things in parallel. Be aware that each phase builds on top of the previous ones, so it is practically impossible to “skip” any steps.

Conclusion

Some companies may know they have costly data quality issues, and some may not. However, everybody prefers good quality data. Hard to argue with that. The question for many is how to reach their data quality happyland. Every company is different. To provide guidance on this journey, the framework we have provided helps identify where you are, and how to take commensurate steps. The order in these steps may not be set in stone, but the underlying dependencies help determine what to do and how to assess progress.

Organizations typically move through awareness creation, raising attention for the data quality problems at hand. Then the next step is developing skills and competencies. This can be training staff, ranging from front-line data-entry to backend data warehouse ETL specialists. But this phase also includes improving user interfaces to enable better data-entry, or supporting data warehouse staff specialist technology (data cleansing tools, etc.). Finally, to make change sustainable, the root causes for data quality problems need to be considered. These typically lie in poorly aligned objectives.

The entire transformation path can be summarized as inform-educate-transform, where each phase builds upon the previous. After data non-quality has been driven out of one process, the organization will learn

to signal similar opportunities in other processes, and the quest for ensuring data quality and making it the default continues.

¹ Jack Olson (2003) Data Quality – the Accuracy Dimension

² Larry English (1999) Improving Data Warehouse and Business Information Quality

³ Ralph Kimball & Joe Caserta (2004) The Data Warehouse ETL Toolkit

⁴ Philip Crosby (1980) Quality is Free

Tom Breur

Tom Breur runs XLNT Consulting, www.xlntconsulting.com, committed to helping companies make more money with their data