



"turning data into dollars"

Tom's Ten Data Tips – April 2008

Data Quality Assessment

Data quality (DQ) assessment is as much about assessing *data* as it is about the impact data quality (or lack thereof) has on business processes. The business case for DQ comes from documenting how data flaws hamper the business. In this information age data is considered an asset that should be managed and leveraged just like any other tangible asset. DQ assessment is a part of auditing to ensure responsible corporate governance.

Thomas Redman has estimated that data quality problems can cost companies as much as 8-12% of revenues. The GMA found that 12-23% of sales and administrative personnel time is spent correcting invoices, orders, purchase orders, etc. 35% of all stock outs, resulting in a loss of sales, are due to data quality issues. Item catalogue errors take on average \$60-\$80 per error to fix and catalogs average an error rate of over 30%. Gartner estimates that 25% of critical data used in large corporations is flawed. TDWI has reported that poor-quality customer data costs US businesses a staggering 611 Billion per year in postage, printing and staff overhead. DQ is big business.

1. Data Are Outdated As Soon As They Enter The Database

Data are only accurate insofar they represent the "real" state of affairs in the world. First of all this assumes that the initial recording of data is an error free process. We all know that not to be the case. But even worse, the minute data are recorded in the database, the world continues its change – without necessarily notifying the database! You can hardly ever rely on changes outside of the database to be *automatically* transferred to updates of records within the database. How up-to-date the information in the database represents the real world, is commonly referred to as the currency of data.

2. Data Conversion Is A Major Source Of DQ Problems

Data are rarely only accumulated "organically". Typically chunks of data are added or appended through data conversion project(s). Such projects are the culprit of some of the nastiest data quality problems.

One reason for this is that data conversion projects are usually performed under considerable time pressure. There is likely to be insufficient time for adequate profiling of the data source. Furthermore, the converted data source has not been designed with the target database in mind, so there need not be a “natural” mapping. Lastly, meta data is often incomplete or entirely missing, so analysts must resort to data profiling, for which there was no time... The end result is typically that imported data tend to comply to ‘different’ (=lower) quality standards than “native” content of the database.

3. Convincing The CEO A DQ Problem Exists Is A Challenge (And A Must)

Many “issues” in a corporation compete for attention from the board. The reason why it is so difficult to get DQ on the agenda is that senior management is often simply unaware of any problems *associated* with DQ. Upstream DQ problems can cause a whole raft of downstream consequences which may never be associated with DQ. The fact that a report for the board takes weeks to create because of DQ problems can stay completely unbeknown to them (and often does). Few professionals, in any organization, are lauded for raising concern about quality issues. Much the same holds for DQ.

The reason why senior management *must* be involved in DQ improvement programs is because rather than only ‘fixing’ issues or systems, the *culture* needs to change, too. And culture is notoriously difficult to change. It can only start at the top. If employees aren’t aware how important data quality is, then you can’t expect them to make this a top priority. And if management doesn’t acknowledge this, it can’t be ‘fixed’, either.

4. DQ Assessment Needs To Be Grounded In (Financial) Numbers

To get buy-in for a DQ improvement program, you need to make a sound business case. This also involves making an assessment of it’s financial impact. Because the relation between (upstream) DQ problems and downstream business processes is rarely obvious, this can be very hard work. Assumptions behind calculations need to be explicated. To complicate matters there is often some ‘political’ component to this work as data cross organizational boundaries.

Getting attention from senior management (see tip#3) goes a whole lot easier with a compelling business case. And if there is *one* dimension every manager, from every company, in every country understands really well, it is dollars (or Euro's, Pounds, or any other currency you use).

5. Establish DQ Benchmarks Across The Company

For all the time that is being spent comparing systems and technology, remarkably little time is spent considering one of the major implementation success factors: data quality. This is all too readily delegated to underlings, despite its importance, and despite substantial investments made in technology.

We advocate benchmark studies *across* the corporation. Research from the Aberdeen Group has established that 80% of best-in-class companies have such benchmarking programs in place, against a mere 28% among laggards. By continuous tracking DQ gets proper attention, and remains on the CIO's radar. Even if two disparate departments cannot be compared side by side (like OLTP versus manual data entry), tracking of their progress still makes sense. And oftentimes other "peer groups" are available within the company, like two data entry departments, different as they might be in many respects.

6. Consider Doing Both Objective *And* Subjective DQ Assessment

Objective DQ assessment is when a set of DQ rules and their outcomes lead to some "score". Ideally this should be represented in a DQ scorecard for periodic reporting. Sets of DQ rules are then fired periodically, and the findings tabulated in the scorecard. Subjective DQ assessment involves interviewing information consumers to determine their opinion and satisfaction with data quality (typically: accuracy) and availability.

To monitor the success and impact of ongoing DQ initiatives, it's worth considering a questionnaire. Cambridge Research Group (CRG) has created the Information Quality Assessment Survey (IQA) for this purpose, which is pretty much the standard tool to use. The advantage of a combined approach of both objective and subjective assessment is that comparisons between groups of information users can be made, organizations can benchmark against best practice peer companies, which can help focus improvement programs.

7. Manual Reconfirmation Is The Royal Road (To DQ)

Manual reconfirmation is when previously entered data is checked against initial documents/information that were used for data entry. You combine DQ rules with manual checks to confront an automated update process with first hand observation of the initial source.

When stratified samples of records (either table or subject oriented) are selected for reconfirmation, automated DQ rules are put to the ultimate test. No matter how sophisticated and well thought through your set of DQ rules, there is still no surrogate for some in-depth data gazing. As smaller and smaller sets of errors are discovered, the DQ rule set gets improved and refined.

By repeating the iterations of manual reconfirmation and continually focusing on records that have been changed, one converges to ever smaller percentages of error. However, each new step also *introduces* new errors. Even when the greatest care is taken, every iteration involving manually re-entering data will introduce new errors, although it is supposed to *fix* the errors... At some point the effort of another iteration is no longer justified, which is the practical optimum. This iterative process uniquely allows for a very reliable *empirical* estimate of the eventual DQ level.

8. Selecting Validation Samples Is Tricky Business

After DQ rules have been developed one needs to sample from affected records to determine type I and type II errors. Sometimes a data quality rule will inadvertently suggest a change where the original record was in fact correct. In other cases errors remain even after all the DQ rules have been fired. This judgment is made on the basis of samples rather than a census (the entire database).

One DQ error is often detected by multiple DQ rules. A missing or invalid primary key will "touch" multiple records that are linked through the key, for instance. One customer's profile is likely to be represented in multiple tables that will all contain an error for this one invalid primary key. This makes sampling of test records quite complicated as the sampling frame is far from obvious. Because the way records from diverse tables are (supposed to be!) interlinked is not straightforward, determining what the "population" is you are drawing from is not at all clear. Furthermore, quality metrics are much more meaningful at the subject-, rather than the record or table level. This complicates sampling considerably. Advanced mathematical skills

are required to determine an efficient and (more or less) unbiased sample.

9. Data Redundancy Drives Quality

Usually, redundancy in databases and in the corporate data ecology is abhorred. The reason for this is not so much wasted disk storage, but rather the risks/problems associated with supposedly identical entries 'growing' out of sync (data inconsistencies). At the corporate level, we are seeking "one version of the truth". If the same object is represented several times, with different values, this results in inconvenient and sometimes costly conflicts.

However, the eternal quest for greater DQ is driven by an incessant urge to iron out any glitches in the data capturing and ETL processes. Investigating each and every "data conflict" is the engine that drives Total Data Quality Management forward. As solutions to existing problems are implemented, new ones emerge, which spur new improvement initiatives, defining new problems, etc. It is in this sense that "errors" should be welcomed and embraced as improvement opportunities and drivers of progress.

10. Ad Hoc Databases Are Priceless

... at least for DQ assessment. Although often treated with contempt for their lack of maintainability, the dozens of ad hoc databases, spreadsheets, and what you have throughout the organization are of tremendous value when assessing data quality. Apparently there is (was) a business need for creating that database, and chances are that the *user* is also the *creator*. This person is likely to be diligent, motivated and pretty detail oriented.

If there is one thing you can be certain about, it is that if there was a business user who saw value in going through the effort to design and build his own ad hoc database, he is very likely to see to it that the quality within it will be *and remain* fit for purpose! These are invaluable benchmarks to "test" supposedly fixed databases against.