Tom Breur
runs the consulting firm XLNT Consulting (www.xlntconsulting.com) dedicated to helping companies make more money with their data. His fields of interest span data mining, analytics, data quality, IT governance, data warehousing, and business models.

Data quality is everyone's business — Managing information quality — Part 2

Tom Breur

Date Received: 11th July 2009

Abstract

In the first of these two papers, we discussed how to design data quality into your data warehouse (DWH) as you are building it. This second paper deals with maintaining a high level of quality after the DWH has gone live. Once a business intelligence solution has been put in place, ongoing data quality needs to be ensured. Data quality maintenance is supported by an appropriate governance structure: the allocation of decision rights and procedures. For errors that have arisen in batches, a data quality project is appropriate. For process causes of data non-quality, a data quality program is more appropriate. Since both kinds of errors often occur side by side, and also to ensure both short-term improvement and sustainable success, in practice a combination of these two approaches is often called for. As organizations progress through subsequent stages on their data quality journey, different measures and actions are required. To make ongoing data quality certain, you progress through information about sources of non-quality and associated organizational costs, training and awareness throughout the organization in conjunction with supporting tools and technology, and alignment and accountabilities that make producing quality the default.

Journal of Direct, Data and Digital Marketing Practice (2009) **11,** 114–123. doi:10.1057/dddmp.2009.21

Keywords: data quality, business intelligence, data warehousing, data models, data quality programs

Introduction

Everybody ‘wants’ data quality. Unfortunately, in many companies this remains largely a ‘motherhood and apple-pie’ issue. We all agree that it’s important, and that something should be done about it. Yet finding volunteers to actually *do* something about it can be painfully difficult.

Data quality costs are often underestimated

Poor data quality costs a lot of money. More than most managers can imagine. Whenever we do data quality audits, the magnitude of — in particular — downstream costs is invariably an unpleasant surprise. In a report from The Data Warehousing Institute issued in 2006, it was estimated that poor-quality customer data costs US businesses a staggering $611bn a year in postage, printing, and staff overhead.[1] Gains from data improvement programs can likewise be huge. Ralph

Tom Breur
XLNT Consulting
Langestraat 8-03,
SE 5038 Tilburg, The Netherlands,
Tel: +31-6-463 468 75
E-mail: tombreur@xlntconsulting.com

Kimball[2] stated that 'In an ongoing warehouse, clean data comes from two processes: entering clean data and cleaning up problems once the data are entered'.

Creating value from data requires a concerted effort throughout the corporation. Given the strategic importance that data plays in competitive markets, this can only be achieved if IT and business find a mutual language, a mode of operation, that overcomes the traditional disconnect we observe in so many corporations.

Do the following dialogues sound familiar? 'Those guys from IT always deliver late, they just can't get their planning right'. 'Those business people can't tell us what they want. In these specs they mistake facts for dimensions, and dimensions for facts. How can we build anything right like this?' I recorded these quotes at a company I was recently working with. Small wonder that their productivity was less than optimal. It looked more like coexistence than cooperation to me.

When we consider data quality in BI solutions, there are two phases. The first phase is about merging data sources, and integrating them in your data warehouse (DWH). We covered this in the previous paper titled 'Data Quality is Everyone's Business — Part 1, designing quality into your data warehouse'.

The second phase is about guaranteeing data quality once your DWH is in place, and installing practices that facilitate making quality the default. This will be the topic of this second paper.

Better IT governance leads to more quality

IT governance should depend on the lifecycle of the company

The way that IT and DWH projects are governed should reflect where a company stands in its lifecycle, and the nature of the markets in which it chooses to compete. But how many companies consciously *choose* their IT governance structure to reflect strategic consideration and acknowledgement of the maturity of markets and needs of business units?

An appropriate IT governance structure needs to be chosen in relation to company strategy and prevailing market conditions. In the early life stages of a company or brand, you will be more concerned with agility and nimbleness than controlling cost. You want to give everyone in the company the opportunity to 'jump' on opportunities they see, and therefore they need freedom to make their own choices, minimally restricted by corporate IT policies.

As the company matures, IT expense as a proportion of revenue is likely to grow too high, so now you will move to a structure that controls cost better. Whereas initially it was fine if multiple solutions were doing much the same thing (as long as people can acquire them *quickly*), now you will try to rationalize your IT portfolio to save on cost.

Governance implies a paradox between control and enabling

Management has been struggling with the long time paradox of encouraging and leveraging the ingenuity of employees while ensuring compliance with the overall vision and principles.[3] We want a work environment that enables creativity, yet at the same time we need rules and guidelines to ensure that a common framework is followed. IT governance means specifying decision rights and accountabilities to encourage desirable behaviour in the use of IT. Good IT governance harmonizes decisions about the management and use of IT with entrepreneurial behaviours and business objectives.

What 'the best' IT governance is depends on the corporate strategy and surrounding market dynamics. When growth is the most important objective, for instance, decisions must be pushed down low. This may result in lower standardization and consequently higher overall investment and maintenance costs. This is because you need to be nimble to quickly meet local needs. The fastest delivering solution isn't always the most economical. Opportunity cost has priority under these conditions and this should offset higher IT investment.

In a mature and saturated market, cost containment is more likely to be a key objective, and, hence, a more centralized model probably works better. Here standardization should drive down maintenance cost, and help leverage procurement power. What is most important is that corporations *choose* their governance model, with an eye out on strategy and prevailing market conditions.

Data quality projects versus programs

In the previous paper, we discussed how DWH development affects data quality. For the initial development, ensuring data quality in large part equals getting your extract, transform, load right.[4] For ongoing quality concerns, data quality follows from carefully aligning IT or business intelligence with internal stakeholders. In this paper, we discuss ongoing efforts to ensure data quality.

We make a distinction between *ad hoc*, one-time improvement initiatives and ongoing data quality improvement programs. The former may be labelled data quality *projects*; the latter we will refer to as data quality *programs*.

Databases are often the result of mergers and acquisitions

Many databases in large corporations have not grown organically, that is, at least not entirely. Either disjoint systems have been merged at some point in the past or the customer portfolio may have grown through merger and acquisitions. For one reason or another, expected data quality levels from different sources tend to differ.

This is often because the data models didn't quite 'match', and some fields were not present in a contributing database. Suppose you are merging two customer databases; one of them holds a field for gender but the other one doesn't. What do you do? Will you drop this field form both databases or will you leave it empty for records from the source system without this attribute? Or yet something else?

Let's suppose you decide to keep this gender field. Now you have a choice to make. If you keep the field and leave it missing for the new records, it is difficult to discern records from the old system (with the field available) that were left empty from the new sister records that never had this field to begin with. At least when you are looking *only* within this column. At the time you do the merge, the distinctions are still pretty crisp. But a database is like an organism: it grows, gets changed and updated; it almost has a life of its own.

A database merge should be preceded by extensive profiling

The best way to merge two databases is to do extensive data profiling to determine all the rules that need to be applied when migrating.[5] In practice, however, these projects are often performed under considerable time pressure, and there simply may not be sufficient time to do thorough profiling. The end result is empty fields, inconsistencies, and 'variable' levels of data quality for different sources. Repairing these problems afterwards is always more time-consuming than getting it right when migrating, but alas, business needs sometimes rightfully prevail. Post hoc fixes, a data quality project, can be a solution here.

Data quality projects

Whether you are trying to 'fix' omissions from earlier data conversions or awareness has risen about the costs of non-quality data, one-time efforts to improve or 'clean up' a database can be a reasonable approach. When original sources (paper or electronically scanned) are still available, manual reconfirmation of data entry is an option.

Data redundancy can be used to derive imputation rules

In large DWHs with many data sources, there may be sufficient redundancy across source systems to compare values for a record across data providers, and derive quality improvement rules from this.[6]

Let's suppose the same data element, in this case 'last name', is available from multiple source systems. Not all systems are created equal, and there may be one that is considered leading. So a business rule says, take field 'last name' from the leading system, for instance a central customer relationship management application. When there is a 'leading system' that serves as the first place to turn to for information, considered 'the truth', IT professionals call that the 'System Of Record' (S-O-R). But what to do if 'last name' is *missing* from the S-O-R?

Since every customer is supposed to have a last name, this field cannot remain empty. You can derive a heuristic or algorithm to deal with the records for which 'last name' is missing. There might be several other systems besides the S-O-R in which 'last name' is held: web forms, customer service, billing, etc. Either one of these systems is leading, again, or they can be combined to determine the 'best guess' for last name, when it is missing from the S-O-R.

'Batch' and 'process' errors are usually mixed

Quite apart from the 'how to' of data quality projects, there are some poor data quality considerations to take into account. It may or may not be very clear how the database errors have arisen. In general, you can expect a mix of *ad hoc* reasons and process causes. The former causes 'batches' of erroneous records to enter; the latter tends to cause much more gradual quality decay. A classic example of an *ad hoc* reason might be a prior database migration.

An example of a process cause for data quality errors would be decaying data-entry quality. Decaying data-entry quality might be caused because this topic was omitted from the induction of a particular cohort of newly trained staff. Consequently, a gradual influx of errors begins to appear as a result of their (poor-quality) work.

In a previous example, we talked about two databases that were joined. For the records from the system without the 'gender' field you introduce a batch of missings, an example of an *ad hoc* cause for errors. When a new employee hasn't been instructed to fill out this field 'gender', for a while you might get missings or invalid values for new customers she signs up. Or maybe some branches or even regions stop recording gender while the majority of offices still do. These would be examples of process causes for poor data quality.

In practice, when you investigate data quality in 'live' databases, you can expect a mix of *ad hoc* and process causes for non-quality. The patterns in errors or missings associated with process causes are stochastic rather than deterministic (like with *ad hoc* causes) and therefore much harder to discern from the data.

Data quality programs

For process causes of data non-quality, a data quality program is more appropriate than a project.[7] As organizations move through successive phases, different efforts are required (Figure 1).

From unconsciously incapable to consciously incapable

When organizations begin their voyage to data quality, they transform from being unconscious of their lack of capability to being conscious of their lack of capability. This move is driven by *information*: information about sources of non-quality, for instance, and the accompanying costs for the business.

Quantifying costs of error helps raise awareness

It is crucially important to *quantify* the costs associated with downstream process breakdown. This is difficult and will require

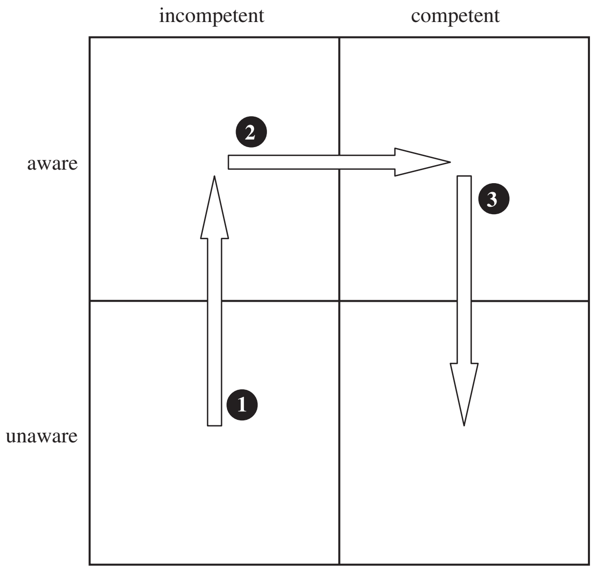

Figure 1: Organizing for data quality

making some (explicit!) assumptions, but by all means provide a quantitative estimate of costs. For some reason, there is one dimension that is understood very well by managers at all levels, anywhere in the world, and it is called 'dollars' (or € or £, or what you have).

When staff repeatedly need to look up customer details in external systems, or an incorrect phone number results in a failed contact attempt, all these small aggravations can add up to substantial loss of efficiency and time for the company. We all get duplicate mailings, and many companies have difficulty in identifying households to suppress duplicate or maybe even conflicting marketing messages. Address deduplication and household identification may not appear the most 'sexy' job, but it is crucially important in delivering a quality experience to your customer.

If no one quantifies the total magnitude of these costs, not only will it be difficult to get management attention, but they will also be in the dark about what kind of effort would be commensurate to actually *do* something about these problems.

In logistics, erroneous data can cause delays, or shipments that get lost altogether. Delays can result in rework and express freight charges. Few good things can happen with inventory in stock. It can get damaged or lost, the due date can expire, and all this time you invest capital in goods that are on their way to the destination. The faster they get there the better. Information about total number of express shipments and resulting aggregate cost helps to raise attention to such problems. The same for damaged goods in inventory, write-offs on perishable goods, etc.

Tools and competence drive the growth to conscious capability

From consciously incapable to consciously capable

As organizations move from incapable to capable, awareness is surpassed by knowledge and skills about how to deal with data quality issues. This phase is characterized by training of staff and implementation of dedicated data quality technology. This could be software for deduplicating name/address records based on fuzzy logic, or monitoring ETL conflicts in an audit dimension, installing an 'Error mart' where bad data are isolated in quarantine, etc.

This transition should apply to both database administration *and* business process owners. To make quality a habit, training should extend all the way from management and senior staff to new employee introduction and developmental training programmes.

With one of our clients we worked on data quality. We had started out with a data quality project that entailed manual reconfirmation of data entry. After all (well, at least hopefully *most*) of the errors had been corrected in the database, a data quality program was initiated.

One of the tools we put in place to support the data quality program was a data quality scorecard. This scorecard tracked accuracy of data entry by having a statistically stratified sample of forms entered more than once. Staff knew about this quality-control mechanism, but could not tell when they would be doing an initial or a duplicate entry. The data entry work they performed was identical. To arrive at an estimate of data entry accuracy, duplicate entries were confronted and the number of conflicts was counted.

Some fields were more costly than others when in error. For this reason, multiple scorecards were derived. During induction training, it was pointed out which fields (potentially!) could be used for the 'high value' scorecard. This was a credit card application form, and in order to preserve the secrecy of the exact scoring algorithm the exact scorecard fields could not be disclosed.

So although the exact composition of the scoring algorithm remained undisclosed to all, it was clear to everyone that an 'overall' and 'scorecard' accuracy metric were derived from the application forms. These data quality scorecards were carefully monitored by senior management.

Data stewards could drill down into scorecard fields, all the way to the individual data entry staff member. In this way, any drop in accuracy could be analysed and traced back to teams or individual team members, specified per group of items or in some cases even individual fields on the application form. A seemingly innocent misspelling of the cardholder's name, for instance, leads to a costly card re-issue and aggravation for the customer. Date of birth is crucial for compliance, etc.

With data stewards in place, and quality control supported by a data quality scorecard, we observed an interesting phenomenon. With *no* additional measures, merely raising attention for the importance of data quality, and constant feedback on error rates, the accuracy kept climbing to levels previously deemed 'impossible'.

More ergonomic screens lead to fewer data entry errors

Another measure we have repeatedly observed to aid quality of data entry is to improve the ergonomics of data entry screens. In our experience, more user-friendly screens that facilitate *faster* data entry also lead to higher accuracy. 'Intelligent' free-form fields and smart drop-down menus can lead to remarkable improvement in error rates.

To summarize the transition from consciously incapable to consciously capable, you train both data entry staff and technical staff to make a concerted effort to drive down error rates. Often, they can be supported by better (or more user-friendly) system interfaces. Sometimes data quality can be supported by special-purpose data quality solutions (home grown or vendor-built) like matching deduplication software or data quality scorecards to monitor ongoing accuracy in detail.

Proper business alignment drives defaulting to quality

From consciously capable to unconsciously capable

The final step from conscious capability to *unconscious* capability requires change in structure and/or accountabilities. This is typically accomplished through organizational development consulting, either internally or externally. The objective is to align business targets so that producing data quality becomes the *default*. If, for instance, data entry staff get rewarded for speed, but not for accuracy, that needs to change.

Conceptually what needs to happen to enable business alignment is bringing together problem holder and problem owner. The problem holder is the person who experiences 'pain' from a problem; a problem owner is the person who controls the resources needed to resolve it.

When the DWH team is faced with conflicting data from disparate sources, they are expected to reconcile these differences nonetheless. In this example, the staff sorting out the ETL logic are the problem holders. System owners of supplying systems in this case are problem owners: they own the data. Maybe structural changes like enterprise application integration or enterprise information integration are needed to resolve lingering data conflicts.

Another classic example is poor-quality data entry. When these data are used for corporate reporting, everybody using these reports is a problem holder. They suffer from the errors in these reports. Management of data entry staff here are the problem owners; *they* can enforce better-quality entry, for instance by rewarding not only speed, but also quality of input.

In all cases, resolutions lie in making the problem owner feel some of the problem holder's 'pain' so he becomes motivated to do something about the problem.

Data quality projects ensure quick wins

Short-term versus long-term fixes

As organizations become aware of the (hidden) costs of data non-quality the question arises how to make improvements. A data quality project may be in order to assure quick wins. One single 'clean up' is not likely to be enough, though, to maintain long-term quality. The results are likely to follow the graph in Figure 2.

If the processes that were in place and that resulted in non-quality remain unchanged, quality will gradually decay to the old level. New (poor-quality) data that get entered after the clean-up will slowly but certainly drive quality down again.

A long-term fix of the problems requires a data quality program to ensure quality becomes the norm for new data entering the system. However, because the program will take some time to become effective

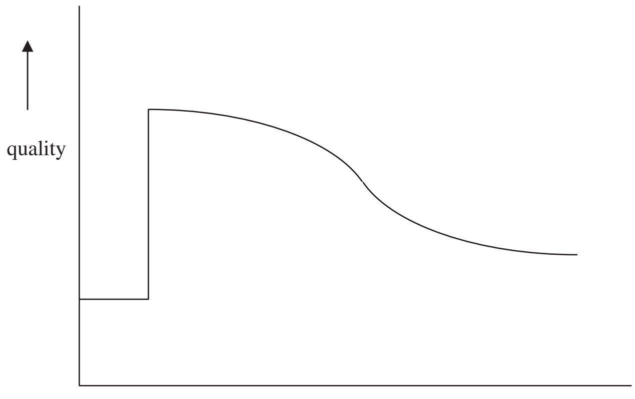

Figure 2: Long-term effects of data quality projects (only)

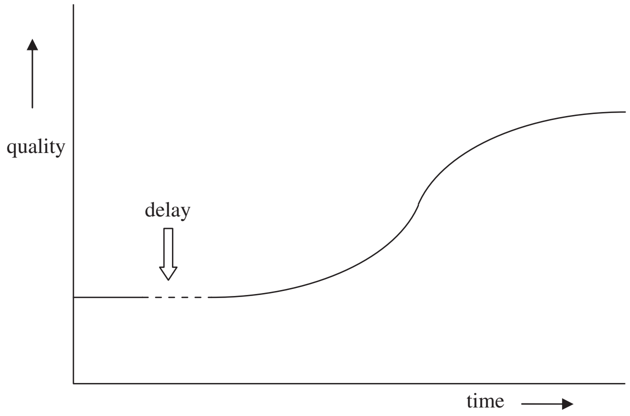

Figure 3: Short-term effects of a data quality program (only)

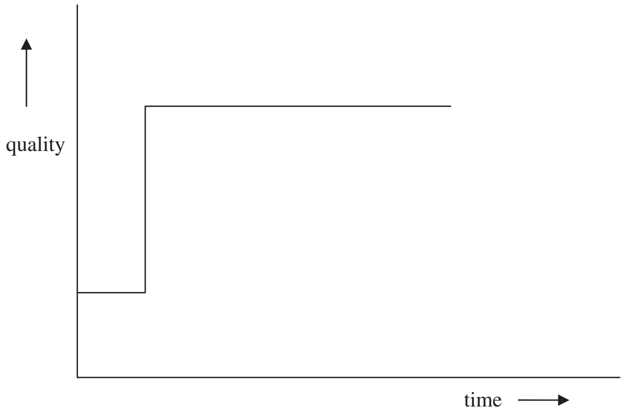

Figure 4: Long-term effects of data quality project and program

and because it will only affect *new* data entering the system, the results are likely to look like the graphic representation in Figure 3.

That is because the old, poor-quality data, will continue to pollute data sources (far out) into the future.

A combination of these two approaches will both ensure rapid improvement and ascertain that the higher-quality standard remains effective (Figure 4).

Now rapid yet sustainable improvement is achieved.

Best practice is to periodically assess data quality by means of a quality scorecard. By monitoring data quality on an ongoing basis, improvement becomes measurable (and thus manageable) and awareness about data quality will grow.

Conclusion

Data quality is, or should be, everyone's concern. For reasons we have outlined, this all-important topic cannot be left to IT or your business intelligence department. Data volumes around us are exploding, and corporations rely on intelligent use of data to create sustainable competitive advantage. This makes extracting the most possible value from (in particular) proprietary data a strategic imperative.

On your data quality journey, you will need to take different measures, different actions, depending on how far you have progressed. Initial stages are characterized by surfacing issues and raising visibility for (hidden) costs associated with data non-quality. After awareness about the consequences of poor-quality data has sufficiently risen, training of staff and selecting tools to support quality need priority. And when attention to data quality as well as the needed skills and tools are in place, it is time to reconsider how to organize work and the company structure in order to ensure that producing quality becomes the default.

Data quality projects can make sense to deal with sources of data non-quality that have originated in batches. For structural improvement of processes, a data quality program makes more sense. Because of the intricate nature of many sources of data quality issues, in practice a combination of the two is most likely called for.

Finally, once the DWH project has been 'finished' successfully, the business is hopefully more receptive to a message that DWH and business intelligence professionals have come to realize for a while already: a DWH initiative is *never* finished. As Kimball[2] states, '… each data warehouse is continuously evolving and dynamic'. Possibly a hard sell, but once embraced, a conclusion that can take business intelligence return on investment to a new level.

References

1. Olson, J. (2003) *Data Quality — The Accuracy Dimension*, Morgan Kaufman, San Francisco.
2. Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W. (1998) *The Data Warehouse Lifecycle Toolkit*, Wiley, NY.
3. Weill, P. and Ross, J. (2004) *IT Governance*, Harvard Business School Press, Boston.
4. Kimball, R. and Caserta, J. (2004) *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*, Wiley, NY.
5. Maydanchik, A. (2007) *Data Quality Assessment*, Technics Publications, Bradley Beach, NJ.
6. Herzog, T. N., Schueren, F. J. and Winkler, W. E. (2007) *Data Quality and Record Linkage Techniques*, Springer, NY.
7. Lee, Y. W., Pipino, L. L., Funk, J. D. and Wang, R. Y. (2006) *Journey to Data Quality*, MIT Press, Cambridge.