



**Tom Breur**

September 2005

## **How Non-Quality Data Can Cost Money**

### **Introduction**

When viewed from a high level, the cost of poor quality data can affect a company's bottom-line in *two* ways. First, there's the cost of scrap and rework, and second, missed opportunities.

An example of scrap and rework costs might be when an agent errs in recording a customer's address details, and consequently a marketing premium is sent to the wrong address. Later, the customer calls to complain. The complaint needs to be handled (extra call center time), the address details then need to be entered a second time (rework), and a second premium needs to be sent. The initial premium is scrapped.

An example of missed opportunity costs might be a credit card that is not granted because the calculated credit score (erroneously) falls below the cut-off score, and the customer is rejected. The opportunity to make a sale is lost, when marketing costs were already incurred.

In this whitepaper, I attempt to supply a comprehensive list of potential data quality costs.

### **Cost Categories of Information Quality**

The costs of data quality can be broken down in 3 categories:

1. *Immediate costs of non-quality data.* This happens when the primary process breaks down as a result of erroneous data. Or, information scrap and rework, when immediately apparent errors or omissions in the data need to be circumvented in support of the primary business process. For example, data entry of a non-valid ZIP code requires back-office staff to look this up again and correct it before sending out a product.
2. *Information quality assessment or inspection costs.* These are costs/efforts expended for (re)assuring processes work properly. Every time a 'suspect' data source is handled, the time spent to seek reassurance of data quality is an irrecoverable expense.

3. *Information quality process improvement and defect prevention costs.*  
Broken business processes need to be improved to eliminate unnecessary information costs. When a data capture or processing operation malfunctions, it requires fixing. This is the long-term investment needed to avoid further losses.

### **1. Immediate costs of non-quality data**

#### Process failure

For example, capturing erroneous customer data like address, contact information, account details.

- Irrecoverable costs; e.g. premiums sent in vain to non-existing customer addresses.
- Liability and exposure costs; for instance credit risk losses when data quality problems cause erroneously offering credit to a customer who is not considered creditworthy on the basis of self-supplied information.
- Recovery costs of unhappy customers; time spent handling complaints.

#### Information Scrap and Rework

- Redundant data handling; because many processes are 'known' to rely on inaccurate data, it is customary for front-line and back-office staff to maintain little private "lists" of all sorts. These serve merely as a backup or improved version of what is available in the primary database. Apart from further problems like 'maintenance' and 'recovery' not being possible for these private lists, such activities are redundant, and non-value adding.
- Costs of chasing missing information; a field that has not been filled out properly, or not at all, needs to be looked up later on in the process. Excess time and costs, inefficiency, and not in the least place an aggravation factor. Time spent looking up missing information is not being spent servicing the customer better.
- Business rework costs; e.g. reissuing a credit card that was sent out with a misspelled customer name.
- Workaround costs; when a primary key is missing or faulty, laborious fuzzy matches need to be performed to match records. This kind of work is challenging, and eats up precious time of the most highly skilled database workers.
- Data verification costs; e.g. costs of reworking data entry. But also, analyses by knowledge workers must begin by checking the correctness of data available before beginning analysis.
- Program rewrite costs; rewriting programs that fail to run because of invalid entries found in the data. E.g.: sometimes pre- or post-conversion scripts needed to be written to deal with the content of source systems prior to loading in a Data Warehouse environment.
- Data cleansing and correction costs; when feeds are processed to load into the Data Warehouse, these data need to be transformed for reasons that stem from quality issues. Any data cleansing and scrubbing that needs to be performed in the ETL process is essentially redundant and

unnecessary insofar this is caused by faulty initial data entry. For example, when a mailing is done on the basis of a problematic customer file, dedicated scripts need to be run to deal with the (known!) errors in the address fields. This process *needs to be repeated for every mailing*. Since such customer files are often shared across departments and systems, *source* changes need to be negotiated with all end users of these data.

- Data cleansing software costs; data cleansing software (like Vality, Ascential, etc.) is usually very expensive. However, there's a tradeoff between scarce labor doing this 'by hand', and the fact that ETL data quality software to help with such tasks typically has very high license costs. Purchase may sometimes prove remarkably economical when related to (often unseen) labor costs for manually improving data quality.

#### Lost and missed opportunity costs

- Lost opportunity costs; when e.g. misspelling customer name on the card causes the customer to not use their card (instead of calling up to complain about this) the business loses their future revenue.
- Missed opportunity costs; when unhappy customers directly influence their social environment, they generate negative publicity. This will make it harder to sell to people in the social network of displeased customers.
- Lost shareholder value; information quality puts a drain on precious resources (scarce database experts), preventing knowledge workers from performing value added work towards market share growth. Scarce human resources are often a bottleneck towards progress, like running one more marketing campaign, delivering insight in a product portfolio's performance, etcetera.

### **2. Information quality assessment or inspection costs**

- People spend time in assessment processes when they are aware of suspect data quality; in any database project, each and every file of questionable quality needs to be inspected for data quality problems first. This time is irreplaceable, forever lost and never recouped in any way. Merely *assessing* if data is of sufficient quality is specialist work. This requires access to scarce resources that are often a bottleneck towards progress.

### **3. Information quality process improvement and defect prevention costs**

- Development costs to rework existing front-end applications; data entry applications need to enforce data quality by performing validity checks, and minimizing keystrokes and eye-hand movements. On the basis of usability findings, interface improvements invariably lead to *both* higher efficiency *and* better data quality.
- Management attention to redefine accountabilities and monitor improved information quality; steering the organization towards higher data quality requires changing accountabilities and continuously monitoring

improvement. This topic will need to stay high on management's agenda to create lasting improvement.

## Conclusion

Problems in data quality often go unnoticed. It can be both a source of process inefficiencies (timeliness), as well as operational costs (direct and indirect losses). In neither of these cases is it apparent that improvement is possible from enhancing data quality.

One of the pernicious consequences of suboptimal data quality is that the *cost* of poor quality data is usually hidden. Lack of data quality is not obvious to those not deliberately looking for it. Quantifying costs isn't always easy. What makes the indirect costs of poor data quality so pernicious is that the relation between data quality problems and its consequences is non-obvious, and often only occurs with a substantial time delay. Therefore, the connection between downstream consequences and poor quality data is often not made, and the problems are not attributed to their true cause.

The *cause* of many downstream data quality costs can easily remain largely hidden (e.g. data quality), and therefore insufficiently subject to management attention and intervention. Also, progress after improvement efforts is gradual, relatively slow, in large part 'cultural', and therefore difficult to monitor and track.

Another, and probably the most significant problem caused by poor-quality information, is that it frustrates the most valuable resource of the company: its employees. Non-quality information prevents knowledge workers from performing their job effectively. On top of that, it alienates customers because of wrong information about them, and to them. Customer data is the raw material that needs to be managed for what it is: a strategic resource.

Data quality is far *more* than accurate data entry. It stems from monitoring downstream data usage, maintaining comprehensive and up-to-date meta data, and nurturing a corporate culture of naturally doing things right at the first attempt. Only then will knowledge workers learn to *expect* data quality, and enforce it because it's the natural thing to do. Letting data quality slide will promote a culture of negligence, and disdain for the use of one's most precious assets: customer information.

The case for accurate source data is further underlined when one realizes that the source in and of itself does little more than support primary processes, which is fine. However, the greater value to the organization comes from *enhancing* these data, from deriving new information from source data.

The investment in improving information quality is recouped several times in decreased costs, and improved value of information to accomplish strategic business goals.

Rapid access to high quality data is *the* decisive factor in an organization's ability to assess and adapt it's business model to changing market conditions. As corporations become ever more 'digitized', those that get a grip on their data quality assurance processes can reap great rewards. In a highly turbulent market this may well be the critical factor in determining the survivors in a competitive business, and therefore prove to be ultimately priceless.

## **Resources**

Larry P. English (1999) Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. Wiley, ISBN 0-471-25383-9

Jack E. Olson (2003) Data Quality: the Accuracy Dimension. Morgan Kaufman, ISBN 1-55860-891-5

Sid Adelman, Larissa Moss & Majid Abai (2005) Data Strategy. Addison-Wesley, ISBN 0-321-24099-5